Journal of Computing and Information Technology - CIT 18, 2010, 3, 245–256
doi:10.2498/cit.1001478

A Framework of Indexation and Document Video Retrieval Based on the Conceptual Graphs

Mounir Zrigui, Mubarak Charhad and Anis Zouaghi

Research Unit of Technologies of Information and Communication (UTIC), Faculty of Science, Monastir University, Tunisia

Most of the video indexing and retrieval systems suffer from the lack of a comprehensive video model capturing the image semantic richness, the conveyed signal information and the spatial relations between visual entities. To remedy such shortcomings, we present in this paper a video model integrating visual semantics, spatial and signal haracterizations. It relies on an expressive representation formalism handling high-level video descriptions and a full-text query framework in an attempt to operate video indexing and retrieval beyond trivial low-level processes, semantic-based keyword annotation and retrieval frameworks.

Keywords: video indexing and retrieval, visual content, conceptual graphs, visual semantics facet, signal facet

1. Introduction

Video data can be modeled based on its visual content (such as color, motion, shape, and intensity) [1], [17] and semantic content in the form of text annotations [6]. Because machine understanding of the video data is still an unsolved research problem, text annotations are usually used to describe the content of video data according to the annotator's understanding and the purpose of that video data.

Although such content descriptions may be biased, incomplete and inaccurate, they still provide much of semantic content that cannot be obtained by current computer vision or voice recognition techniques.

In the case of video, there are a number of specificities due to its multimedia aspect. For instance, a given concept (person, object. . .) can be present in different ways: it can be seen, it can be heard, it can be talked of, and combination of these representations can also occur. Of course, these distinctions are important for the user. A query concerning X as "Show me a picture of X." or as "I want to know what Y has said about X." are likely to give quite different answers.

The first one would look for X in the image track while the second would look in the audio track for a segment in which Y is the speaker and X is mentioned in the transcription.

Also, among all possible relations that could be represented in conceptual graphs, some are especially appropriate for content-based video indexing.

Based on the traditional keywords approach to video information retrieval is that usually too many or erroneous results are returned. To overcome this inconvenience, a solution is to include conceptual descriptions when the documents are indexed. Let us consider, for example, the following query: "documents dealing with the president of the United States". Using the currently existing search techniques like keywords approach, we extract some words as for example "president", "United States" and search the collection of documents with a query that combines these keywords. The problem arising here is that there is no way to specify that the expected answer to this query is a person, and thus several wrong results can be obtained in return of such a query.

Our objective is therefore to propose a model for content representation semantics of video documents allowing the consideration of information from each of the modalities (image, text, sound) and implement this model within a system of indexing and search by content of video documents.

2. Related Works

As far as indexing and retrieval techniques for the visual content are concerned, the first systems (content-based) [12] propose a set of methods based on low-level features such as colors, textures... fully automatic, and able to process queries quickly. Several frameworks dealing with the automatic extraction of the image semantic content have been proposed [7], [14], [11]. Their main disadvantage relies however on the specification of restrained and fixed sets of semantic classes. Indeed, regarding the fact that several artificial objects have high degrees of variability with respect to signal properties such as color and texture variations, an interesting solution is to extend the extracted visual semantics with signal characterizations in order to enrich the image indexing vocabulary and query language. Therefore, a new generation of systems integrating semantics and signal descriptions has emerged and the first solutions [15], [19] are based on the association of textual annotations with relevance feedback (RF). Prototypes such as iFind [15] offer loosely-coupled solutions based on textual annotations to characterize semantics and on a RF scheme operating on low-level signal features. These approaches present two major drawbacks: first, they lack to exhibit a single framework unifying signal features and semantics, which penalizes the performance of the system in terms of retrieval efficiency and quality. Then, regarding the query process, the user is to query both textually in order to express high-level concepts and through several and time-consuming RF loops to complement his initial query. Therefore, it is impossible for such systems to process complex queries combining semantic and signal concepts as well as relations linking them such as “find Bill Clinton and a striped red, white and blue flag behind him”.

All previous works in visual case can’t describe the video content in much depth. The solution consists of combining signal and symbolic characterizations in order to reinforce the semantic gap and to take into account the generic and multifacets descriptions for video content indexing and retrieval.

3. A Multi-facetted Framework for Video Retrieval

As for textual document, video has a specific organization. A video is composed of scenes, each of which describes an event.

A video scene itself is composed of a number of shots, each of which is an unbroken image sequence captured continuously by the same camera. A shot could legitimately be compared to a word, as they are both the basic entities structuring respectively a video sequence and a text fragment.

The visual content of a video shot can be represented by its key frames. Key-frames are images providing a compact representation of the video shots. They can serve as pointers to the given portion of the video content for indexing and retrieval process. Generally, the key frame extraction process is integrated with the processes of shot segmentation. Each time a new shot is identified, the key-frame extraction process is invoked, using parameters already computed during shot boundary detection (SBD) [17]. These parameters are related to visual data such as color, or camera motions descriptors. Key frame selection can differ depending on the application needs. For example, we require only a few key frames (1-2) for a video-captured meeting since camera motions are sparse [17], [5]. Whereas, in a broadcast news document, we find more animation (related to visual aspects) so we have more shots and key-frames. In our approach, the key-frame is selected as the most stable image of a given shot (i.e. image that appears many times at the same shot).

We propose the outline of an image model combining a set of interpretations, each considered as a particular facet of an image, to build the most exhaustive image description. The image is therefore seen as a multi-facetted object with the two principal facets being the physical (considering an image as a matrix of pixels) and the symbolic facets. The symbolic facet, grouping all aspects of the image content and its general context, is itself an aggregation of two basic facets:

— The visual semantics facet describes the image semantic content and is based on labelling image objects with a visual semantic concept.

— The signal facet describes the image signal content in terms of symbolic perceptive features and consists of characterizing image objects with signal concepts. It is itself divided into two sub-facets. The *color subfacet* features the image signal content in terms of symbolic colors. E.g., the image object (USA flag) is associated with symbolic colors *Blue*, *Red* and *White*. The *texture subfacet* describes the signal content in terms of symbolic texture features.

At the core of the image model is the notion of image object (IO), abstract structure representing a visual entity within an image. Its specification is an attempt to operate visual indexing and retrieval operations beyond simple low-level processes or object-based techniques [14] since Ios convey the visual semantics and signal information.

3.1. Representation Formalism

In order to instantiate this model as an image retrieval framework, we need representation formalism capable of representing image objects as well as the visual semantics, spatial and signal information they convey. Moreover, this representation formalism should make it easy to visualize the information related to an image. It should therefore combine expressivity and a user-friendly representation. As a matter of fact, a graph-based representation and particularly conceptual graphs (CGs) [15] are an efficient solution to describe an image and characterize its components. The asset of this knowledge representation formalism is its flexible adaptation to the symbolic approach of image retrieval [16], [8]. It allows indeed to uniformly represent components of our architecture and to develop expressive and efficient index and query frameworks.

Formally, a CG is a finite, bipartite, connex and oriented graph. It features 2 types of nodes: the first one between brackets in our CG alphanumerical representation (i.e. as coded in our framework) is tagged by a concept, however, the second between parentheses is tagged by a conceptual relation. E.g., the CG:

$$[ICTA] \leftarrow (Name) \leftarrow [Conference] \\ \rightarrow (Location) \rightarrow [Tunisia]$$

is interpreted as: the ICTA conference is held in Tunisia.

Concepts and conceptual relations are organized within a lattice structure partially ordered by the IS-A ($\leq$) relation. E.g., Person $\leq$ Man denotes that the concept Man is a specialization of the concept Person, and will therefore appear in the offspring of the latter within the lattice organizing these concepts. Within the scope of the model, CGs is used to represent the video shot content in the logical facet.

The indexing module provides a representation of a video shot document in the corpus with respect to the multi-facetted image model. It is a CG called document index graph.

Also, as far as the retrieval module is concerned, a user full text query is translated into a video shot conceptual representation: the video shot query graph corresponding to the multi-facetted shot description.

After presenting our representation formalism, we now focus on the visual semantics facet and propose conceptual structures handling semantics. We then specify its CG representation.

3.2. The Visual Content Modeling

3.2.1. The Visual Semantics Facet

a) Automatic extraction of semantic concepts

Visual semantic concepts are learned and then automatically extracted given a visual ontology. Several experimental studies presented in [20] have led to the specification of twenty categories or picture scenes describing the image content at a global level. Web-based image search engines (Google, AltaVista) are queried by textual keywords corresponding to these picture scenes and 100 images are gathered for each query. These images are used to establish a list of semantic concepts characterizing objects that can

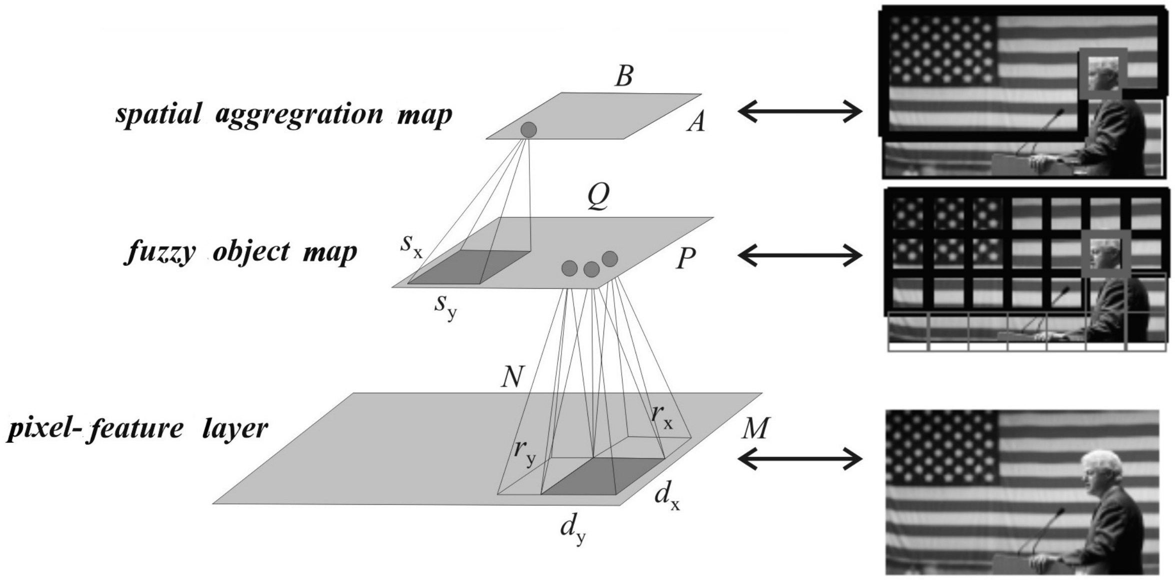

Figure 1. Architecture for visual semantics extraction: The tiled key-frame (1) is subjected to visual token recognition and recognition (2) results are then spatially aggregated to determine semantic concepts (3).

be encountered in these scenes. This list (in particular concepts related to individual names) is enriched with concepts provided by Video Annex [11] and a total of 72 visual semantic concepts to be learned and automatically extracted are specified.

Figure 1 presents the architecture for automatic extraction of visual semantic concepts: a 3-layer feed-forward neural network with dynamic node creation capabilities is used to learn these concepts from 1000 labelled key-frame patches cropped from the training and annotation corpus of the TRECVID 2003 search task. Color and texture features are computed for each training region as an input vector for the neural network. Once the network has learned the visual vocabulary, the approach subjects a tiled key-frame to be indexed to multi-scale, view-based recognition against these visual semantic concepts. A key-frame to be processed is scanned with windows of several scales within the *Fuzzy Object Map* in Figure 1. Each one represents a visual token characterized by a feature vector constructed with respect to the feature vectors of visual semantic concepts exhibited previously. Recognition results are then reconciled across multiple resolutions and aggregated according to configurable spatial tessellation within the *Spatial Aggregation Map*.

b) Model of the visual semantics facet

Image objects are represented by *Io* concepts and visual semantic concepts are organized within a multi-layered lattice ordered by a specific/generic partial order (we propose a part of the lattice in Figure 2). An instance of the visual semantics facet is represented by a set of canonical CGs, each one containing an *Io* type linked through the conceptual relation *vsc* to a visual semantic concept. The basic graph controlling the generation of all visual semantic facet graphs is:

$$[Io] \rightarrow (vsc) \rightarrow [VSC]$$

E.g., the representation of the visual semantics

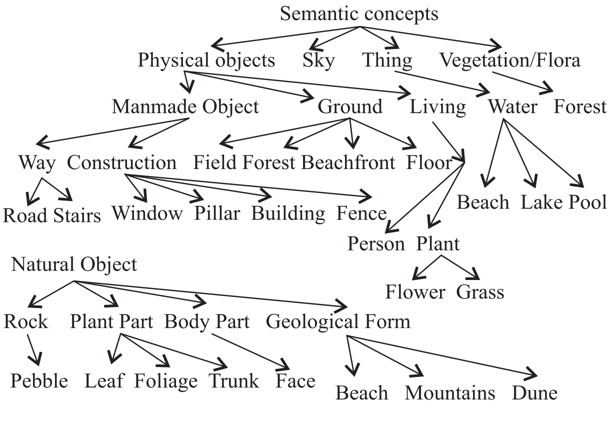

Figure 2. Lattice organizing semantic concepts.

facet for our example image in Figure 1 is

$$[Io1] \rightarrow (vsc) \rightarrow [Clinton]$$

and

$$[Io2] \rightarrow (vsc) \rightarrow [flag]$$

translated as the first IO represents Clinton and the second IO a flag.

3.2.2. The Signal Facet

The integration of signal information within the conceptual level is crucial since it enriches the indexing framework and expands the query language with the possibility to query over both semantics and visual information. After presenting our formalism, we will now focus on the signal facet and deal with theoretical implications of integrating signal features within our multifaceted conceptual model. This integration is not straightforward as we need to characterize low-level signal features at the conceptual level, and therefore specify a rich framework for conceptual signal indexing and querying.

We first propose conceptual structures for the color and texture subfacets and then thoroughly specify their CG representation

3.2.2.1. The Color Subfacet

Integrating signal features within a high-level conceptual framework is not straightforward. The first step is to specify conceptual signal data which correspond to low-level features, therefore specifying a correspondence process between color names and color stimuli. Our symbolic representation of color information is guided by the research carried out in color naming and categorization [2] stressing a step of correspondence between color names and their stimuli. We will consider the existence of a formal system Snc of color categorization and naming which specifies a set of color words Cat with a cardinal Ccat. These color words are the Ci. Within the scope of this paper, 11 color words:

$$C1 = red,\ C2 = white,\ C3 = blue,$$
$$C4 = grey,\ C5 = cyan,\ C6 = green,$$
$$C7 = yellow,\ C8 = purple,\ C9 = black,$$
$$C10 = skin,\ C11 = orange$$

spotlighted in [29] are described in the HVC perceptually uniform space by a union of brightness, tonality and saturation intervals.

a) Conceptual specification

Each IO is indexed by a color index concepts ($CICs$) feature the color distribution of image objects by a conjunction of color words and their corresponding integer pixel percentages. The second image object (Io2) corresponding to the semantic concept *flag* in Figure 1 is characterized by the CIC:

$$< r : 40, w : 45, b : 15, g : 0 \ldots >$$

interpreted as Io2 having 40% of red, 45% of white and 15% of blue.

CICs are elements of partially ordered lattices, organized with respect to the query operator processes: either a Boolean or a quantification operator (*at most, at least, mostly, few*) explicited in [21] [22]. Index color graphs link an *Io* type through the conceptual relation *has_color* to a CIC:

$$[Io] \rightarrow (has\_color) \rightarrow [CIC].$$

b) Automatic generation of color subfacet CGs

Here is the algorithm summarizing the automatic generation of all conceptual structures of the color subfacet:

- Given an IO
- compute the RGB value of each of its pixels
- Map it to tonality, brightness & saturation values in the HVC perceptive space
- Determine the associated color word considering the HVC perceptive color word partition [9]
- Store for each color word the percentage of associated pixels
- Generate the associated CIC and the alphanumerical color CG:

$$[IO] \rightarrow (has\_color) \rightarrow [CIC]$$

E.g., the representation of the color subfacet for our example image in Section 1 is:

$$[Io2] \rightarrow (has\_color)$$
$$\rightarrow [< r : 40, w : 45, b : 15, g : 0 \ldots >]$$

TW	B	C	D	I	L	M	N	S	S_p	U	W
%	83.7	85.2	88.9	91.9	94.5	89	86.8	83.4	90	97.3	81.4

Table 1. Cross-validation percentages.

translated as the second image object (Io2) is associated with the CIC :

$$< r : 40, w : 45, b : 15, g : 0 \ldots >$$

(i.e. 40% of red, 45% of white & 15% of blue).

3.2.2.2. The Texture Sub Facet

The study of texture in computer vision has led to the development of several computational models for texture analysis used in several CBIR architectures [7]. However, these texture extraction frameworks mostly fail to capture aspects related to human perception. Therefore, we propose a solution inspired by the work in [12] specifying a computational framework for texture extraction which is the closest approximation of the human visual system. The action of the visual cortex, where an object is decomposed into several primitives by the filtering of cortical neurons sensitive to several frequencies and orientations of the stimuli, is simulated by a bank of Gabor filters. However, as opposed to their work operating at a global level of an image, we will focus on computational texture extraction at the object level. We will therefore characterize each image object by its Gabor energy distribution within seven spatial frequencies covering the whole spectral domain and seven angular orientations. Each image object is then represented by a 49-dimensions vector, with each dimension corresponding to Gabor energy.

a) Conceptual texture characterization

Although several studies have been proposed for the analysis of the characteristic texture, few proposals were made for the symbolic recognition of this feature. Our symbolic representation of texture is based on research in naming and categorizing textures proposed by Bhushan [3]. We consider the following concepts of texture as a representation of each of these categories:

bumpy, cracked, disordered, interlaced, lined, marbled, netlike, smeared, spotted, uniform and whirly.

These 11 high-level texture words, foundation of our framework for texture symbolic characterization are automatically mapped to the 49-dimensions vectors of Gabor energies through support vector machines [20]. We adopt the one-against-rest approach where a separate classifier is designed for each of the eleven texture words for reasons of optimized inter-class separation. We also associate a confidence value for the classification defined. For this, we use the distance from an IO i to be characterized with texture word t to the decision boundary $ft(i)$ (where ft is the trained discriminate function on the one-against-rest classification problem involving texture word t) and map it on posterior probabilities of recognition. In order to achieve this mapping, we use a 1D logistic classifier which maximizes the likelihood of the classified training IOs. For each of the eleven texture words, the best cross-validation rate is given in Table 1. Let us note that the SVMs are able to label new instances of unknown textures with corresponding texture words with a high accuracy, cross-validation percentages being higher than 80%.

b) Conceptual specification

Each IO is indexed by a texture concept (TC). A TC is supported by a vector structure $\boldsymbol{t}$ with eleven elements corresponding to texture words ***twi***. Values $\boldsymbol{t}[i]$, $i \in \{1, \ldots, 11\}$ are booleans stressing that the texture distribution of the considered IO is characterized by the texture word twi.

E.g., the second IO ($Io2$) corresponding to the semantic concept *flag* in Figure 2 is characterized by the TC:

$$< B : 0 \ldots D : 0 \ldots L : 1 \ldots U : 1 \ldots >$$

translated as $Io2$ being characterized by the texture word **striped** (lined). TCs are elements of partially ordered lattices which are organized respectively to the type of the query processed. The basic graphs controlling the generation of all texture subfacet graphs link an

Io type through the conceptual relation tx to a texture concept:

$$[Io] \rightarrow (has\_tx) \rightarrow [TC]$$

c) Automatic generation of texture subfacet CGs

Here is the algorithm summarizing the automatic generation of all conceptual structures of the color subfacet:

- Given an IO
- Compute its associated 49-dimensions vector of Gabor energies
- Map it to the linked texture words through the explicated SVM architecture
- Compute the posterior recognition probabilities of association
- Generate the associated TC & the texture CG:
$$[Io] \rightarrow (has\_tx) \rightarrow [TC]$$

E.g., the representation of the texture subfacet for our example image in Figure 1 is

$$[Io2] \rightarrow (has\_tex) \\ \rightarrow [< B : 0 \ldots D : 0 \ldots L : 1 \ldots U : 1 \ldots >]$$

translated as the second image object ($Io2$, representing the semantic concept *flag*) is associated with the TC $< B : 0 \ldots D : 0 \ldots L : 1 \ldots U : 1 \ldots >$ (i.e. striped/lined).

3.2.2.3. The Spatial Subfacet

In order to model spatial data, we first consider a subset of the topological relations explicited in the RCC-8 theory [5]; 4 relations which are exhaustive and relevant for image querying are chosen. Considering 2 image objects ($Io1$ and $Io2$), these relations are:

— $s_1 = (\mathbf{C}, Io1, Io2)$: $Io1$ covers/is in front of $Io2$ (therefore $Io2$ is behind $Io1$).

— $s_2 = (\mathbf{P}, Io1, Io2)$: $Io1$ is a part of $Io2$.

— $s_3 = (\mathbf{T}, so1, so2)$: $Io1$ touches $Io2$ (externally connected or overlaps).

— $s_4 = (\mathbf{D}, so1, so2)$: $Io1$ is disconnected with $Io2$.

Directional relations Right ($s_5 = \mathbf{R}$), Left ($s_6 = \mathbf{L}$), Above ($s_7 = \mathbf{A}$), Below ($s_8 = \mathbf{B}$) are invariant to basic geometrical transformations (translation, scaling).

Two relations specified in the metric space are based on the distances between image objects. They are the Near ($s_9 = \mathbf{N}$) and Far ($s_{10} = \mathbf{F}$) relations.

An instance of the spatial facet is represented by a set of CGs, each one containing 2 *Io* types linked through the previously defined spatial relations (more details are found in [18], [23]).

An IO is characterized by its centre of gravity io_g as well as two pixel sets: its interior, noted io_i and its boundary, noted io_b. To deal with the automatic computation of topological relations [9], two image objects $Io1$ and $Io2$ are characterized by intersections of their interior and boundary sets:

$$io1\_i \cap io2\_i,\ io1\_i \cap io2\_b,\ io1\_b \cap io2\_i \\ \text{and } io1\_b \cap io2\_b.$$

Each topological relation is mapped to the results of these intersections, e.g. $(DC, so1, so2)$ iff.

$$io1\_i \cap io2\_i = \emptyset,\ io1\_i \cap io2\_b = \emptyset, \\ io1\_b \cap io2\_i = \emptyset \text{ and } io1\_b \cap io2\_b = \emptyset.$$

The interest of this computation method relies on the association of topological relations to the previous set of necessary and sufficient conditions involving attributes of spatial objects (i.e. interior and boundary). The computation of directional relations between $Io1$ and $Io2$ relies on the relative position of their centers of gravity.

Finally, to distinguish between near and far relations, we use the D_{nf} constant given by

$$D_{nf} = d(\vec{0}, 0.5 * [\sigma_1, \sigma_2]^T)$$

where d is the Euclidean distance between the null vector $\vec{0}$ and $[\sigma_1, \sigma_2]^{\mathrm{T}}$ is the vector of standard deviations of the localization of centers of gravity for each IO in each dimension from the overall spatial distribution of all IOs in the corpus. D_{nf} is therefore a measure of the spread of the distribution of centers of gravity of IOs. This distance agrees with results from psychophysics and can be interpreted as: the bigger the spread, the larger the distances between centers of gravity are. We will say that two IOs are **near** if the Euclidean distance between their centers of gravity is inferior to D_{nf}, **far** otherwise.

a) Conceptual specification

Each pair of IOs are related through a spatial concept (SpC), compact structure summarizing spatial relationships between these IOs. A SpC is supported by a vector structure **sp** with ten elements corresponding to the previously explicited spatial relations. Values ***sp[i], i***$\in${***1,...,10***} are booleans stressing that the spatial relation si links the two considered IOs. E.g., Io1 and Io2 are related through the SpC:

$$< C : 1,\ P : 0 \ldots N : 1,\ F : 0 >$$

translated as $Io1$ covering and being near to $Io2$ ($Io2$ being therefore behind $Io1$). SpCs are elements of a partially ordered lattice explicited in [21]. The basic graph controlling the generation of all spatial subfacet graphs links two Io types through the conceptual relations $agent\_1$ and $agent\_2$ to a SpC:

$$[Io1] \leftarrow (agent\_1) \leftarrow [SIC] \rightarrow (agent\_2) \rightarrow [Io2].$$

b) Automatic generation of spatial subfacet CGs

Here is the algorithm summarizing the automatic generation of all conceptual structures of the spatial subfacet:

- Given a pair of IOs, $Io1$ and $Io2$
- Associate a topological relation to the results of interior and boundary sets of $Io1$ and $Io2$
- Compare the centers of gravity of both IOs to determine the directional relations linking them
- Compute $dEuc(Io1\_g, Io2\_g)$ and compare it to D_{nf} to determine the near /far relations between $Io1$ and $Io2$
- Generate the associated SIC and the alphanumerical spatial CG:

$$[Io1] \leftarrow (agent\_1) \leftarrow [SIC] \rightarrow (agent\_2) \rightarrow [Io2].$$

E.g., the representation of the spatial subfacet for our example image in Figure 1 is $[io1] \leftarrow (agent\_1) \leftarrow [< c : 1, p : 0 \ldots n : 1, f : 0 >] \rightarrow (agent\_2) \rightarrow [io2]$, translated as $io1$ covering and being near to $io2$.

4. The Query Module

Our conceptual architecture is based on a unified full-text framework allowing a user to query over the visual layer.

This obviously optimizes user interaction since the user is in 'charge' of the query process by making his information needs explicit to the system. The retrieval process using CGs relies on the fact that a query is also expressed under the form of a CG. The representation of a user query in our model is, like image index representations, obtained through the combination (joint operation) of CGs over all the facets of visual layers. Without going into details, a simple grammar composed of a list of the previously introduced visual concepts, as well as the specified visual relations is automatically translated into an alphanumerical CG structure.

We distinguish several categories of requests:

— Request Event: The application event involves actions (often real) in the document. For example, the query: "find the video segments showing a rally". Representation with the formalism of conceptual graph of this query is given.

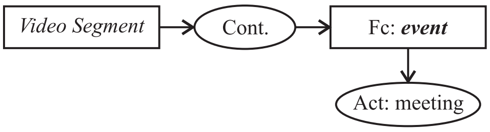

Figure 3. Example query event.

— Temporal request: The temporal request integrates constraints from the temporal aspect. For example, the query: "find the video segments showing both events ending at the same time". Representation with the formalism of conceptual graph of this query is given in the following figure:

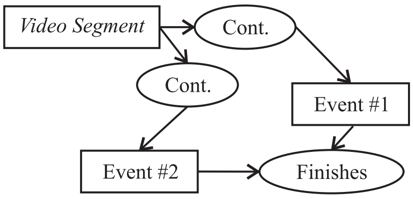

Figure 4. Example query temporal.

— Audio semantic request: The audio semantic request combines the descriptions connected with the audio contents. For example, the request: "find the video segments in which Bill Clinton spoke". Representation with the formalism of conceptual graph of this query is given in the following figure:

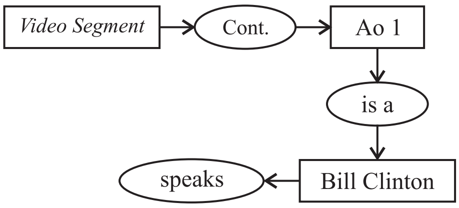

Figure 5. Example query semantics audio.

— Visual semantic request: The visual semantic request combines the descriptions connected with the visual contents. For example, the request: "find the video segments showing Bill Clinton". Representation with the formalism of conceptual graph of this query is given in the following figure:

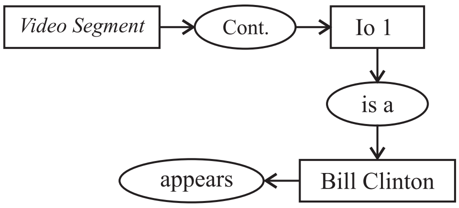

Figure 6. Example of visual semantic query.

— Semantic request signal: The semantic request signal can search the video based on low-level descriptions. For example, the request: "Find the video segments showing a flag with a red, white and blue striped texture".

$[Flag] \rightarrow (has\_color) \rightarrow [< C1 : 1, C2 : 1, C3 : 1, \ldots, C9 : 0 >].$
$\rightarrow (has\_texture) \rightarrow [< T1 : 0, T2 : 0, T3 : 0, \rightarrow T4 : 0, T5 : 1, \ldots >]$

— Multimodal request: The multimodal request allows searching the video based on a description from several modalities (image, audio or text). For example, the request: "Find the video segments showing Bill Clinton speak on Iraq and where at least part of the American flag is visible". Representation with the formalism of conceptual graph of this query is given in the following figure:

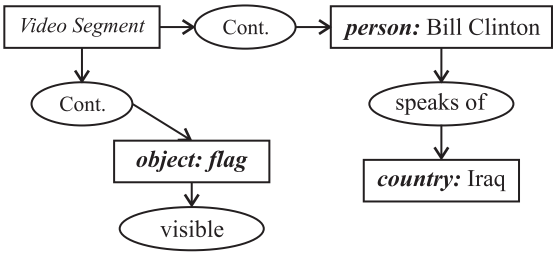

Figure 7. Example of multimodal semantic query.

5. The Matching Process

The correspondence between a document d (d may be a video or a segment of video document) and a query q is determined using the operator of projection of the graph G_q representing the request in the graph G_d representing the document. There is a projection of the graph G_q on a graph G_d if there is one under graph G'_d of G_d which is a restriction of G_q.

Given a query q and document d, there may be zero, one or more such projections. We denote $\prod(q, d)$ all of these projections.

When many documents match the query in accordance with such correspondence, it is necessary to order them. For this, we calculate the relevance for each video shot VS (document). It is calculated by combining the measures called for exhaustivity and specificity. This correlation model is based on an extension of the logic model of VanRijsbergen logical model proposed in [13]. We use a function F to combine exhaustivity and specificity measures:

$$Relevance(VS, Q) = F[E(VS \rightarrow Q), S(Q \rightarrow VS)]$$

Exhaustivity quantifies to which extent the video shot satisfies the query. It is given by the value of $E(VS \rightarrow Q)$, E being the exhaustivity function. ***Specificity*** measures the importance of the query themes within the considered video shot, it is given by the value of $S(Q \rightarrow VS)$, S being the specificity function.

The function F values are to be proportional to the values of the exhaustivity and specificity functions: it takes its values in the [0,1].

$$F(a, b) = 0 \textit{ if } a = 0 \textit{ or } b = 0;$$

$$F(a, b) = 1 \textit{ if } a = 1 \textit{ and } b = 1;$$

We have chosen the trivial multiplication operation.

$$F(a, b) = a * b;$$

6. Application: TRECVID Topic Search

Our prototype called CLOVIS[1] implements the theoretical framework exposed in this paper and validation experiments are carried out on the TRECVID 2004 corpus comprising 128 videos segmented in 48817 shots, each one itself represented by a key-frame.

The search task is based on *topic retrieval* where a topic is defined as a formated description of an information need, therefore involving multiple characterizations. The complexity inherent in topic search revolves around the difficulty to design the intended meaning and interrelationships between the various characterizations. We therefore design the evaluation task in the context of manual search, where a human[1]expert in the search system interface is able to interpret a topic and propose an optimal query to be processed by the system. 24 multimedia topics developed by NIST for the search task express the need for video concerning people, things, events, locations... and combinations of the former. The topics are designed to reflect many of the various sorts of queries real users propose: requests for video with specific people or people types, specific objects or instances of object types, specific activities or locations or instances of activity or location types. We compare our system with the mainstream TRECVID 2004 systems operating manual search [10].

Carnegie Mellon University proposes a manual search system using text retrieval based on ASR and closed captioning to find candidate shots and then re-ranking the candidates by linearly combining scores from multimodal features or re-ranking weights trained by logistic regression.

The manual search system of the *National University of Singapore* is based on a generic query analysis module. They use 6 query-specific models and the fusion of multi-modality features such as text, OCR, visual concepts... their work being inspired by text-based question-answering techniques.

Finally, the *Lowlands team (CWI Amsterdam & Twente University)* proposed a generative probabilistic model for video retrieval based on dynamic (at the shot level), static (at the key-frame level) and audio/language (using ASR) characterizations. Their queries are created by manual construction and selection of visual examples.

The Recall/Precision curves in Figure 8 illustrate the average results in terms of mean average precision obtained for the 24 multimedia topics. The average precision of CLOVIS (0.0868) is approximately 14.2%, 19% and 22.3% higher over respectively the average precisions of the CMU (0.076), Lowlands (0.073) and NUS (0.071) systems. This clearly indicates that, on average, the first video shots returned by our system are particularly relevant compared to the first video shots retrieved by other systems.

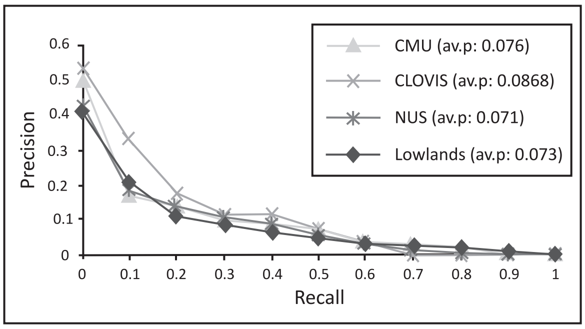

Figure 8. Recall/Precision curves for TRECVID topics.

The obtained results allow us to state that when performing topic search and therefore dealing with elaborate queries which combine multiple sources of information (here visual and audio semantics, signal features) and thus require a higher level of abstraction, the use of an "intelligent" and expressive representation formalism (here the CG formalism within our framework) is crucial. As a matter of fact, our framework outperforms state-of-the-art TRECVID 2004

[1] CLOVIS: Conceptual Layer Organization for Video and Indexing and Search.

systems by proposing a unified full-text framework optimizing user interaction and allowing querying with precision over visual and audio/speech descriptions.

7. Conclusion

We proposed the specification of a framework combining semantics, signal and spatial characterizations within a strongly-integrated architecture to achieve greater retrieval accuracy. We introduced image objects, abstract structures representing visual entities in order to operate video indexing and retrieval operations at a higher abstraction level than state-of-the-art frameworks. We specified the multiple facets, their conceptual representation and finally proposed a full-text unified and rich query framework.

Our experimental contribution consists of the (partial) implementation of the CLOVIS prototype. We have integrated the proposed model in the video indexing and retrieval system by content in order to evaluate its contributions in terms of effectiveness and precision. Experimental results on a TRECVID corpus allowed us to validate our approach which aims to solve several problems such as:

— Exploit semantic descriptions of content

— Facilitate handling and access to large video databases based on the description level symbolic

— To gather descriptions from different submedia in the same schema modelling.

In further work, we propose in the short term to exploit the results of visual analysis (signal) and integrate different representations at the level of the modeling, then complete the integration of model in a video search system for assessing the contribution of the proposed model on another corpus. On the long run, we propose to use external knowledge to enrich the descriptions in the schema modeling.

References

[1] G. AMATO & AL., An Approach to a Content-based Retrieval of Multimedia Data. *Multimedia Tools Appl.* 7, pp. 9–36., 1998.

[2] B. BERLIN, P. KAY, *Basic Color Terms: Their Universality and Evolution*. UC Press, 1991.

[3] N. BHUSHAN, & AL., The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. *Cognitive Science* 21(2), pp. 219–246, 1997.

[4] J. S. BORECZKY, L. A. ROWE, Comparison of video shot boundary detection technique. In *IS&T/SPIE*, USA, February 1996.

[5] E. ETIEVENT & AL., Assisted Video Sequences Indexing: Motion Analysis Based on Interest Points. *ICIAP*, 27–29, 1999.

[6] R. FABLET, P. BOUTHEMY, Statistical motion-based video indexing and retrieval. *Conf. on Content-based Multimedia Information Acces*, Vol. 1, pp. 602–619, Paris, France, 2000.

[7] S. FENG, & AL., Multiple Bernoulli Relevance Models for Image and Video Annotation. *CVPR*, pp. 1002–1009, 2004.

[8] J. L. GAUVAIN & AL., The LIMSI Broadcast News Transcription System. *Speech Communication* 37, pp. 89–108, 2002.

[9] Y. GONG & AL., Image Indexing and Retrieval Based on Color Histograms. *Multimedia Tools and App. II*, pp. 133–156, 1996.

[10] W. KRAAIJ, A. SMEATON & P. OVER, TRECVID 2004 – An Overview.

[11] J. H. LIM, Explicit query formulation with visual keywords. *ACM MM*, pp. 407–412, 2000.

[12] C.-Y. LIN & AL., VideoAnnEx: IBM MPEG- 7 Annotation Tool for Multimedia Indexing and Concept Learning. *IEEE on ICME*, 2003.

[13] Y. LU & AL., A unified framework for semantics and feature based RF in image retrieval systems. *ACM MM*, pp. 31–37, 2000.

[14] J. NIE, An outline of a General Model for Information Retrieval Systems. *SIGIR* (1998) pp. 495–506.

[15] I. OUNIS AND M. PASCA, RELIEF: Combining expressiveness and rapidity into a single system. *SIGIR*, pp. 266–274, 1998.

[16] J. F. SOWA, *Conceptual structures: information processing in mind and machine*. Addison-Wesley, 1984.

[17] G. QUNÉOT, Shot Boundary Detection Task: CLIPS System Description and Evaluation. In *TREC*, Gaithersburg, Maryland, pp. 13–16, Nov. 2001.

[18] V. VAPNIK, *Statistical Learning Theory*. Wiley, NYC 1998.

[19] J. YANG & AL., Finding Person X: Correlating Names with Visual Appearances. *CIVR*, pp. 21–23, 2004.

[20] X. S. ZHOU & T. S. HUANG, Unifying Keywords and Visual Contents in Image Retrieval. *IEEE Multimedia* 9(2), pp. 23–33, 2002.

[21] M. CHARHAD, Modèles des documents video bases sur le formalisme des GCs pour l'indexation par le contenu Sémantique. PhD Thesis, UJF, Grenoble France, 2005.

[22] M. CHARHAD, M. ZRIGUI & G. QUÉNOT, Une approche conceptuelle pour la modélisation et la structuration sémantique des documents vidéos. *SETIT*, 2005.

[23] M. CHARHAD, M. ZRIGUI & H. AFLI, A Framework Integrating Signal / Semantic Visual Characterizations for Conceptual Video Retrieval. *International Conference on Artificial Intelligence and Pattern Recognition (AIPR-09)*, Orlando, Florida, USA, July 13–16, pp. 325–332, 2009.

Received: April, 2009
Revised: January, 2010
Accepted: March, 2010

Contact addresses:
Mounir Zrigui
Mubarak Charhad
Anis Zouaghi
Monastir University, Faculty of Science
Research Unit of Technologies of
Information and Communication (UTIC), Tunisia
e-mail: mounir.zrigui@fsm.rnu.tn

MOUNIR ZRIGUI received his PhD from the Paul Sabatier University, Toulouse, France in 1987 and his HDR in computer science from the Stendhal University, Grenoble, France in 2008. Since 1986, he was a Computer Science Professor at Brest University, France, and later at the Faculty of Science of Monastir, Tunisia. He started his research, focused on all aspects of automatic processing of natural language (written and oral), in RIADI laboratory and continued it in UTIC Laboratory (Tunisia). He has run many research projects and has published many research papers in reputed international journals/conferences.

MUBARAK CHARHAD received his PhD degree in computer science from the University of Grenoble (France), in 2005. Currently, he is an Associate Professor at the Computer Science Department, University of Gabès, Tunisia. His research areas include artificial life and pattern recognition.

ANIS ZOUAGHI received his PhD from the Manouba University, Tunisia in 2008 and Master's degree from the INPG, Grenoble, France in 2000. Since 2009, he has been a Computer Science Professor at Gabès University, Tunisia. He started his research, focused on the processing of the Arabic language, in RIADI laboratory and continued it in UTIC Laboratory.