Towards a High-Level Audio Framework for Video Retrieval Combining Conceptual Descriptions and Fully-Automated Processes

Mbarek Charhad and Mohammed Belkhatir

IMAG-CNRS, BP 53, 38041 Grenoble Cedex 9, France
{charhad, belkhatm}@imag.fr

Abstract. The growing need for 'intelligent' video retrieval systems leads to new architectures combining multiple characterizations of the video content that rely on highly expressive frameworks while providing fully-automated indexing and retrieval processes. As a matter of fact, addressing the problem of combining modalities within expressive frameworks for video indexing and retrieval is of huge importance and the only solution for achieving significant retrieval performance. This paper presents a multi-facetted conceptual framework integrating multiple characterizations of the audio content for automatic video retrieval. It relies on an expressive representation formalism handling high-level audio descriptions of a video document and a full-text query framework in an attempt to operate video indexing and retrieval on audio features beyond state-of-the-art architectures operating on low-level features and keyword-annotation frameworks. Experiments on the multimedia topic search task of the TRECVID 2004 evaluation campaign validate our proposal.

1 Introduction

The size, heterogeneity of content, and the temporal characteristics of video data pose many interesting challenges to the video indexing and retrieval community. Among these challenges is the modeling task for effective content-based indexing and user access capabilities such as querying, retrieval and browsing.

Video data can be modeled based on its visual content (such as color, texture, shape, motion…) [1], [17] audio content [7]. Models such as VideoText [9], VSTROM [2] and VideoGraph [16] whether they use the video annotation (stratification) approach or the keyword-based annotation approach [3], [4] to represent video semantics, fail to model semantic relationships among the concepts expressed in the video. The importance of capturing video semantic associations lies in the fact that it can greatly improve the effectiveness of video querying by providing knowledge-based query processing [13], [4]. This is due to the fact that human beings always have multiple expressions or terms for the same or similar semantics. For example, “sport” and “baseball” do not match syntactically but match conceptually. Furthermore, video semantic associations can be used for flexible, knowledge-based video browsing. Existing visual content-based video browsing approaches are mostly static.

Other techniques for video modeling analyze the semantic content by considering object hierarchies. For example, the model proposed in [1] allows hierarchical

Y.-S. Ho and H.J. Kim (Eds.): PCM 2005, Part I, LNCS 3767, pp. 820 – 830, 2005.

abstraction of video expressions representing scenes and events. It provides the possibility to assign multiple interpretations to a given video segment and functionalities for creating video presentations.

However, all tasks involving automated visual characterization at the key-frame level involves heavy computational treatments for low-level extraction while providing very poor recognition results. Indeed, the results of fully-visual manual runs at the TRECVID 2004 evaluation campaign do not go beyond a 0,01 average precision rate [11]. As for now, dealing with fully-automated visual characterization appear to penalize greatly video indexing and retrieval frameworks both as far as the retrieval results are concerned and the computational load involved for low-level feature extraction. As a consequence, text annotations are usually used to describe the video content according to the annotator's expertise and the purpose of this content [9]. However, such descriptions are biased, incomplete and often inaccurate since subjected to the annotator's point-of-view. Also, manual annotation is a costly task which cannot keep pace with the ever-growing size of video collections.

We therefore strongly believe that being able to model aspects related to the audio content is crucial in order to assist a human user in the tasks of querying and browsing. Also, since users are more skilled in defining their information needs using language-based descriptors [14], this modeling task is to consider a symbolic representation of audio features as textual descriptors. Indeed, a user would naturally formulate his desire to being provided with videos where Y is speaking about X with the full-text query " Show me videos where Y is speaking about X.". For this, the traditional keyword-based approaches in state-of-the-art video architectures would appear clearly not satisfactory since they fail to take into account aspects related to conceptual and relational descriptions of the video content. Indeed, we believe that in order to process a query such as "Show videos displaying Bill Clinton speaking" (proposed in the framework of the TRECVID 2004 topic search track), a system is to characterize concepts such as Bill Clinton and relations such as the fact that he is speaking.

In order to process these queries involving non-trivial information needs, we propose in this paper to integrate audio descriptions within a unified full-text framework by considering:

- The specification of a rich video model featuring all characterizations of the audio content. It is based on audio objects, structures abstracting the audio flow related to a video document (they are detailed in section 2).
- The integration of a knowledge representation formalism (conceptual graphs) in order to instantiate the video audio model within a video indexing and retrieval framework and therefore specify indexing, querying and matching processes.
- The specification of fully-automated processes to build and manipulate the conceptual index and query descriptions. Indeed, the strength of our approach relies in the specification of high-level descriptions of the video content while being able to process video corpus of relatively important size.
- The evaluation of our theoretical proposition in the framework of the multimedia topic search track of the TRECVID 2004 evaluation campaign. We will show that it outperforms state-of-the-art systems which do not take into account conceptual and relational characterizations of the video audio content.

In the remainder, we detail in section 2 our conceptual audio framework for video indexing and retrieval. Section 3 deals with the automatic conceptual characterization

of audio/speech features and section 4 tackles the index and query conceptual representations. We finally discuss validation experiments conducted on the TRECVID 2004 corpus in section 5.

2 A Bi-facetted Model for Audio Characterization

We propose the outline of an audio model for video indexing and retrieval supported by an expressive knowledge representation formalism. It describes the information related to video shots through audio segment flows and is represented by a bi-facetted structure (fig.1):

- The audio object facet characterizes audio objects (AOs), abstraction of audio elements extracted from the audio flow.
- The audio semantic concept facet provides the semantic description related to each AO. It is based on concepts such as person identity, organization, location… and consists in specifying the speaker identity in each shot as well as the characterization of the audio content being spoken of.

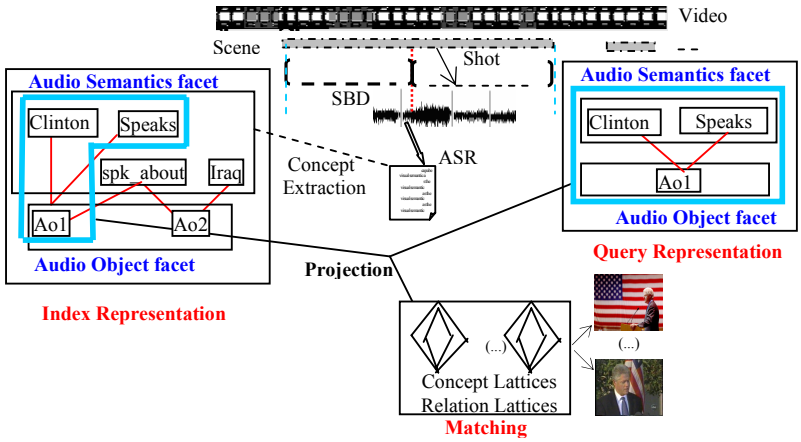

Fig. 1. Video Model and General Description of Indexing, Querying and Matching Processes

In order to instantiate this model as a video retrieval framework, we need a representation formalism capable of representing audio objects as well as the audio semantics they convey. Moreover, this representation formalism should make it easy to visualize the information related to the video shot. 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 characterize the audio content of a video shot. The asset of this knowledge representation formalism is its flexible adaptation to the symbolic approach of multimedia retrieval [14]. 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, related 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:[PCM_05] → (Name) → [Conference] → (Location) → [Jeju] is interpreted as: the PCM 2005 conference is held in Jeju.

Concepts and conceptual relations are organized within a lattice structure partially ordered by the IS-A (≤) relation. E.g., Person ≤ 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 are used to represent the audio content of a video shot within index and query stuctures.

The indexing module provides the audio representation of a video shot document in the corpus with respect to the explicited model. It is itself a CG called video shot document audio index graph. In fig. 1, a video shot belonging to the corpus is characterized by a bi-facetted conceptual audio representation.

Also, as far as the retrieval module is concerned, a user full-text query is translated into a video shot conceptual audio representation: the video shot query audio graph corresponding to the bi-facetted audio description. In fig. 1, the query "Find shots of Bill Clinton speaking" is translated into a bi-facetted audio conceptual representation.

The video shot query audio graph is then compared to all audio conceptual representations of video shot documents in the corpus. Lattices organizing audio semantic concepts are processed and a relevance value, estimating the degree of similarity between video shot query and index audio CGs is computed in order to rank all video shot documents relevant to a query.

After presenting our formalism, we now focus on the characterization of the audio content by proposing its conceptual specification and the automatic processes for its generation.

3 Automatic Conceptual Characterization

There are several approaches in the literature for audio segmentation based on speaker change detection. Approaches proposed in [10], [12] assume that the probability of a speaker change is higher around silence regions and use speech-silence detectors to identify the speaker change locations.

Other applications related to audio content characterization through automatic speech recognition (ASR) provide the best segmentation results. The LIMSI system [5] has indeed a 93 % recognition success rate in the TRECVID evaluation task.

ASR consists in analyzing the audio flow for transcribing speech. The output of such process is a structured textual information with temporal descriptions. In our proposition, we use results of speaker-based segmentation and ASR generated in the TRECVID 2004 collection. We aim at analyzing the transcription content for speaker identification in each segment using linguistic patterns. First, we propose to categorize these patterns. Then, we apply them to detect speaker's identities.

We target broadcast news as specific audiovisual content. This kind of documents presents some particularities:

- The number of speakers in this document is limited to three: presenter, reporter and intervening individual.
- The transition between two speakers is often provided except in the case of presenters introducing themselves only at the beginning of news. These expressions may be classified in two groups: those referring to speakers in the next, current or previous segments and those allowing the speakers to identify themselves. These expressions appear just before or after the identity of the person mentioned. All these expressions are called linguistic patterns.

3.1 Speaker Identity Characterization

There are two challenging tasks for video semantic content indexing using audio flow. The first one is specifying who the speaker is in each video segment. As we previously mentioned it, we use linguistic patterns for speaker identity detection. We propose three categories of patterns:

- The first category is for detecting the identity of the speaker who is speaking in the current segment. For example, when the speaker introduces himself: "... this is C.N.N. news I' m [Name]... "
- The second category is for detecting the identity of the speaker who has just spoken in the previous segment.
- The third category is for detecting the identity of the person who will speak (speaker of the following segment).

Table 1 summarizes some of the patterns that we use in our approach gathered by category.

Table 1. List of some linguistic patterns

Previous segment	Current segment	Next segment
thank you...	I'm [name]	tonight with [name]
[name]	[name] CNN	ABC's [name]
thanks… [name]	[name] ABC	[name] reports
[name] reporting	…….	……
……		

The detection process consists in parsing each segment and identifying passages containing one of these patterns. We then apply a tool for identity recognition. For this, we use a named-entity extraction tool based on two lists of concepts. The first contains a set of first names (~12400) and the second contains common words except family names. We compare neighboring words of each detected pattern with the content of the 2 lists. If neighboring words are for example elements of the first list and not present in the second list, we can estimate that they deal with a person's identity. We then infer the corresponding identity by comparing its localization with respect to a linguistic pattern.

Our approach is summarized in fig. 2 which displays the complete process for automatic speaker identification. We tested our approach on the TRECVID 2004 collection and obtained a success rate of 82 % for automatic speaker's identity recognition.

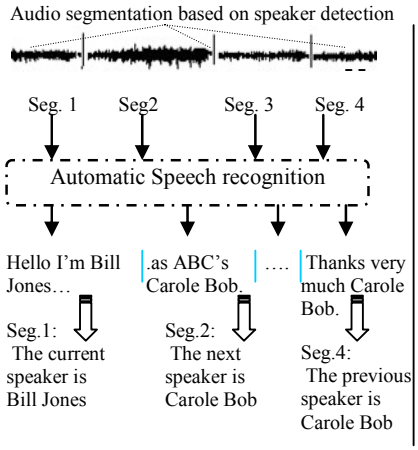

Fig. 2. Identity detection approach: overview

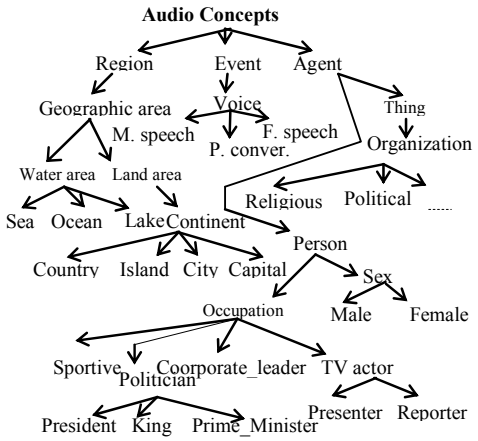

Fig. 3. A part of the audio concept lattice

3.2 Conceptual Content Characterization

To extract concepts from audio content, we parse transcriptions files to detect symbolic information by comparing their items to elements in the four concept classes (e.g. person identity, the name of a city, an organization, etc.). We then extract, from each document, the concepts which correspond to each class. This process is based on the projection of each document on the audio concept lattice (partly displayed in fig. 3). We exploit linguistics forms to specify concepts such as the expressions Mr., Mrs. appearing before a person identity or propositions like "in", "at" and "from" before a localization concept (place). Here is the algorithm summarizing the concept extraction process:

Given an audio segment

Extract AOs by projecting the transcription of the audio segment on specific ontologies

- Collect AOs belonging to the concept category person
- Collect AOs belonging to the concept category place
- Collect AOs belonging to the concept category organization
- If AOs belong to the concept category person, specify AOs corresponding to speakers using linguistic patterns.

After presenting the automatic conceptual characterization, we now focus on its conceptual instantiation within a video indexing and retrieval framework. We moreover detail the generation of index and query structures.

4 Conceptual Instantiation Within Index and Query Structures

4.1 CG Representation of the Audio Semantics Facet

Extracted concepts are related through specific audio relations to audio objects (AOs). Considering for example two audio objects (Ao1 and Ao2), these relations are spk(Ao1) where Ao1 belongs to the concept class person translated as Ao1 speaks and spk_abt(Ao1,Ao2) translated as Ao1 is about Ao2. For the audio layer modeling we use all automatic extracted concepts, they are labeled audio semantic concepts (ASC).

Audio objects are represented by Ao concepts. They are linked to ASCs by the CG: [Ao]→(asc)→[ASC]. The audio characterization of a video shot is provided by a set of canonical CGs: [Ao1]→(spk_abt)→[Ao2] and [Ao1] →(spk).

4.2 Index CGs

The index conceptual representation of a video shot is a CG obtained through the combination (**join** operation [14]) of canonical CGs over the audio facet. For our example video shot of fig.1, the unified conceptual representation is:

JOIN[[Ao1]→(asc)→[Clinton] ∩ [Ao2]→(asc)→[Iraq] ∩ [Ao1]→(spk) ∩ [Ao1]→(spk_abt)→[Ao2]]

4.3 Query Module

Our conceptual architecture is based on a unified full-text framework allowing a user to query over the audio layers. 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 index representations, obtained through the combination (join operation) of CGs over all the facets of audio layers. Without going into details, a simple grammar composed of a list of the previously introduced audio concepts, as well as the specified audio relations is automatically translated into an alphanumerical CG structure. For instance, the query string: "Bill Clinton speaking" is translated into the joint unified graph: JOIN[[Ao1]→(asc)→[Clinton] ∩ [Ao1]→(spk)].

The Matching Process. The matching framework is based on an extension of VanRijsbergen's logical model proposed in [18]. We define the relevance of a video shot VS with respect to a query Q as a function of the exhaustivity measure which quantifies to which extent the video shot satisfies the query:

$$\text{Relevance}(VS,Q) = P(VS \rightarrow Q)$$

The exhaustivity function P consists in two operations. It first checks that all elements described within the query graph are also elements of the index graph. For this, we use the CG projection operator to compare video shot query and index graphs. This operator allows to identify within the video shot index graph i all sub-graphs with the same structure as the query graph q, with nodes being possibly restricted, i.e.

they are specializations of q nodes. Πq(i) is the set of all possible projections of query graph q into video shot index graph i. Let us note that several projections of a query graph within an index graph may exist. Then, for each selected video shot, we provide an estimation of its relevance with respect to the query, which corresponds to the quantitative evaluation of their similarity. It is given by the exhaustivity value between query graph q and video shot index graph i:

$$EV(q,i) = MAX\Pi q(i) \left[\Sigma_{ASCq \text{ concept of } q,\ ASCi \text{ matching concept of } i} IA(ASCi)+Cpt\_Match\ (ASCq,ASCi)\right]$$

- The IA function measures the importance of an audio semantic concept and is related to the number of times it is pronounced during the corresponding audio segment.
- The Cpt_Match function is the negative Kullback-Leibler divergence between the probabilities of audio query concepts which are themselves certain (i.e. P(ASCq) equal 1) and the posterior recognition probabilities of audio semantic concepts of graph i.

Let us note that brute-force implementations of the projection operator would result in exponential execution times. Therefore, based on the work in [14], we use an adaptation of the inverted file approach for video retrieval. We specify indeed lookup tables associating audio semantic concepts to the set of image index representations that contain these concepts.

5 Application: TRECVID Topic Search

The CLOVIS[1] prototype 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.

AOs are automatically assigned audio semantic concepts as presented in section 5. Finally, all audio alphanumerical representations of CGs linking AOs to audio concepts and relations are automatically generated as presented in sections 3 and 4.

The search task is based on topic retrieval where a topic is defined as a formatted description of an information need text. We therefore design the evaluation task in the context of manual search, where a human expert in the search system interface is able to interpret a topic and propose an optimal query to be processed by the system. Ten multimedia topics and their ground truths 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 CLOVIS with the mainstream TRECVID 2004 systems operating manual search on audio features. The National Taiwan University system is based on ASR tokens and high level features (concepts) using WordNet for word-word distances. Approach proposed in this system aligns the ASR word tokens to the corre-

[1] **C**onceptual **L**ayer **O**rganization for **V**ideo **I**ndexing and **Search.**

sponding shots by calculating their word-to-word distance with the high-level feature COMF tokens. Without complex algorithms and plenty of computing time, this method does lead to an improvement of the performance of the video information retrieval. The Indiana University system named viewfinder uses only text search based on ASR output. Each query is manually formulated as speech keywords and supported by tf/idf term weighting. Their queries are created by manual construction and selection of visual examples [11].

We propose the top retrieval results for 4 multimedia topics in fig. 4. Average precisions results for each of the 10 topics are provided in fig. 5. The mean average precision over the 10 topics of CLOVIS (0.08019) is approximately 45.54% and 49.05% higher over respectively the mean average precisions of the IU (0.0538), and NTU (0.0551) systems.

The obtained results allow us to state that when performing topic search and therefore dealing with elaborate queries involving conceptual and relational audio characterization 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 systems by proposing a unified full-text framework optimizing user interaction and allowing to query with precision over audio/speech descriptions.

Fig. 4. Top 4 retrieval results for topics 133, 130, 135 and 136

In the table below, they are proposed with their translation in terms of relevant textual query terms as input to CLOVIS.

Table 2. TRECVID 2004 Topics and CLOVIS transcription

TRECVID topic	CLOVIS transcription
128.US Congressman Henry Hyde's face, whole or part, from any angle	Henry Hyde speaking or being spoken of
129.US Capitol dome	Washington, White House spoken of
130.Hockey rink	Hockey or N.H.L spoken of
133.Saddam Hussein	Saddam Hussein speaking or being spoken of
134.Boris Yeltsin	Boris Yeltsin speaking or being spoken of
135. Sam Donaldson's face. No other people visible with him	Sam Donaldson speaking or being spoken of
136.Person hitting a golf ball	P.G.A. spoken of
137.Benjamin Netanyahu	Benjamin Netanyahu speaking or being spoken of
142.Tennis player	A.T.P. spoken of
143.Bill Clinton speaking	Bill Clinton speaking

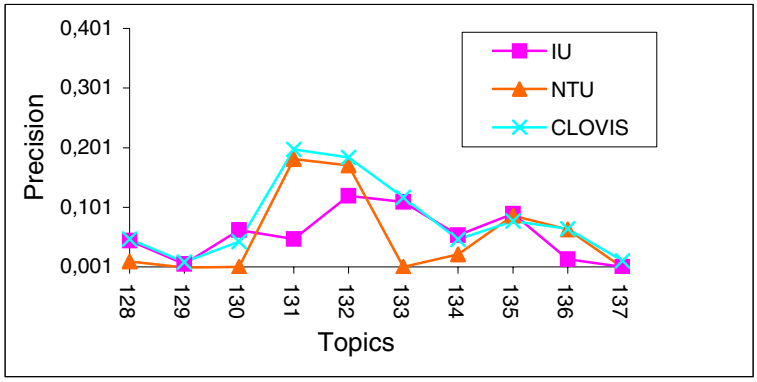

Fig. 5. Average Precision for each of the 10 multimedia TRECVID 2004 topics

6 Conclusion

We proposed the specification of a framework featuring audio characterizations within an integrated architecture to achieve greater retrieval accuracy. We introduced audio objects, abstract structures characterizing the audio content related to a video shot 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, the

conceptual representation of index and query structures and finally proposed a unified full-text query framework. Experimental results on the TRECVID 2004 multimedia topic search task allowed us to validate our approach

References

1. Amato G., Mainetto G., Savino P.: "An Approach to a Content-Based Retrieval of Multimedia Data", Multimedia Tools and Applications 7, 9-36, 1998
2. Arslan U., Dönderler M-E, Saykol E, Ulusoy Ö, Güdükbay U: "A Semi-Automatic Semantic Annotation Tool for Video Databases", Workshop on Multimedia Semantics (SOFSEM'02), The Czech Republic, pp. 1-10, 2002
3. Assfalg J., Bertini M., Colombo C.and Del Bimbo A.: "Semantic Annotation of Sports Videos". IEEE MultiMedia 9(2), 52-60, 2002
4. Bertini M, Del Bimbo A., Nunziati W.: "Annotation and Retrieval of Structured Video Documents", in Proc. of Advances in Information Retrieval, ECIR 2003, Pisa, Italy, 14-16, 2003.
5. Gauvain J-L., Lamel L., Adda G.: "The LIMSI Broadcast News transcription system". Speech Communication 37, 89-108, 2002
6. Gong Y., Chua HC, Guo XY: "Image Indexing and Retrieval Based on Color Histograms". Multimedia Tools and App. II, 133-156, 1996
7. Jiang H.,Danilo Montesi D., Ahmed K. Elmagarmid A. k.: "Integrated video and text for content-based access to video databases". Multimedia Tools and Applications, 1999
8. Jiang H.,Abdelsalam Helal A., Ahmed K. Elmagarmid A. k., Joshi A.: "Scene change detection techniques for video database systems". ACM Multimedia Systems 6, 186–195, 1998
9. Jiang H.,Danilo Montesi D., Ahmed K. Elmagarmid A. k.: "VideoText database systems". Int'l Conf. on Multimedia Computing and Systems, 334–351, 1997
10. Kemp, T., Schmidt, M., Westphal, M., Waibel, A.: "Strategies for Automatic Segmentation of Audio Data". ICASSP, 1423-1426, 2000
11. Kraaij, W. & Smeaton, A. & Over P. "TRECVID 2004– An Overview"
12. Kwon S. & Narayanan S.: "Speaker Change Detection Using a New Weighted Distance Measure". ICSLP, 16-20, 2002
13. Lozano R. and Martin H.: Querying virtual videos using path and temporal expressions. ACM Symposium on Applied Computing, 1998.
14. Ounis, I. & Pasca, M. "RELIEF: Combining expressiveness and rapidity into a single system". SIGIR, 266-274, 1998
15. Sowa, J.F. "Conceptual structures: information processing in mind and machine". Addison-Wesley, 1984
16. Tran D. A., Hua K. A., Vu K.: "VideoGraph: A Graphical Object-based Model for Representing and Querying Video Data". ICCM, 383-396, 2000
17. Quénot, G. "TREC-10 Shot Boundary Detection Task: CLIPS System Description and Evaluation". TREC 2001.
18. VanRijsbergen, C.J. "A Non-Classical Logic for Information Retrieval". Comput. J. 29(6), 481-485, 1986