The VIDI-Video semantic video search engine

Marco Bertini, Università di Firenze, Italy, Marco Rendina, Fondazione Rinascimento Digitale, Italy[128]

Introduction

Video is becoming vital to society and economy. It plays a key role in information distribution and access, and it is also becoming the natural form of communication on the Internet and via mobile devices. The massive increase in digital audiovisual information will pose high demands on advanced storage and retrieval engines, and it is certain that consumers and professionals will need advanced storage and search technologies for the management of large-scale video assets.

Current search engines, however, mostly rely on keyword-based access that uses manually annotated metadata, and do not allow for content-based search of images or videos. At present, even state-of-the-art video search engines are able to annotate automatically only a limited set of semantic concepts, and retrieval is usually allowed using only a keyword-based approach based on a lexicon.

The VIDI-Video project, funded in the 6th Framework Program by the EU, has taken on the challenge of creating substantially enhanced semantic access to video. The project has aimed to integrate and develop state-of-the-art components from many technologies — such as machine learning, audio event detection, video processing, visual feature processing, knowledge modeling and management, interaction and visualization — into a fully implemented audiovisual search engine, combining large numbers of audiovisual concepts and exploiting the interclass similarities of these concepts, as well as using information from different sources: metadata, keyword annotations, audiovisual data, speech, and explicit knowledge.

The international consortium that has worked on the project presents excellent expertise and resources in all these technologies:

- the machine learning with active 1-class classifiers, to minimize the need for annotated examples, was lead by the University of Surrey (UK)
- video stream processing was lead by CERTH (Greece)
- audio event detection was lead by INESC-ID (Portugal)
- visual image processing was lead by the University of Amsterdam (Netherlands)
- interaction and knowledge management was lead by the University of Florence (Italy)
- software consolidation was lead by CVC (Spain)
- provision of data, and evaluation and dissemination was lead by Beeld & Geluid (Netherlands), and FRD (Italy), as application stakeholders.

VIDI-Video has boosted the performance of audiovisual searching by forming a thesaurus of more than 1,000 detectors for the corresponding semantic concepts in the audio, video or combined streams of data. This large thesaurus of detectors can be viewed as the core of a dictionary for video. The elements in such a thesaurus, individually or in combination, provide a semantic understanding of the audiovisual content. In order to reach this goal of semantic understanding, VIDI-Video has improved the state-of-the-art on machine learning techniques, visual and audio analysis techniques and interactive search methods.

The approach followed has been to let the system learn many, mostly weak, semantic detectors instead of modeling a few of them carefully. These detectors describe different aspects of the video content. In combination they create a rich basis for interactive access to the video library. The VIDI-Video system has achieved the highest performance in the

128 The authors can be contacted at: Marco Bertini bertini@dsi.unifi.it; Marco Rendina mrendina@gmail.com

most important object and concept recognition international contests (PASCAL VOC and TRECVid).

The outcome of the project is an audiovisual search engine, consisting of two parts: an automatic annotation part, that runs off-line and processes the videos, computing the audiovisual features, that then are fed to automatic classifiers that create the annotations. In this part there are components for video processing, visual analysis, audio analysis, and learning integrated feature detectors. The automatic annotation part of the system performs audio and video segmentation, speech recognition, speaker clustering and semantic concepts detection. And it can be further expanded, increasing the number of automatic concept detectors, training the system with positive and negative examples of audiovisual concepts.

The second part is the interactive part of the system, and it provides a video search engine for both technical and non-technical users. The interfaces can be run as standalone applications or as web-based search engines that exploit technologies such as ontologies, developed for the semantic web. The system also uses different interfaces to allow different query modalities: free-text, natural language, graphical composition of concepts using Boolean and temporal relations, and query by visual example.

In addition, the ontology structure is exploited to encode semantic relations between concepts, permitting, for example, expanding queries to synonyms and concept specializations.

All subsystems are delivered and available both as standalone and integrated into the complete annotation systems (Figure 1). The modularity and, at the same time, the standalone status of each system warrants developmental independence, and an efficient exploitation.

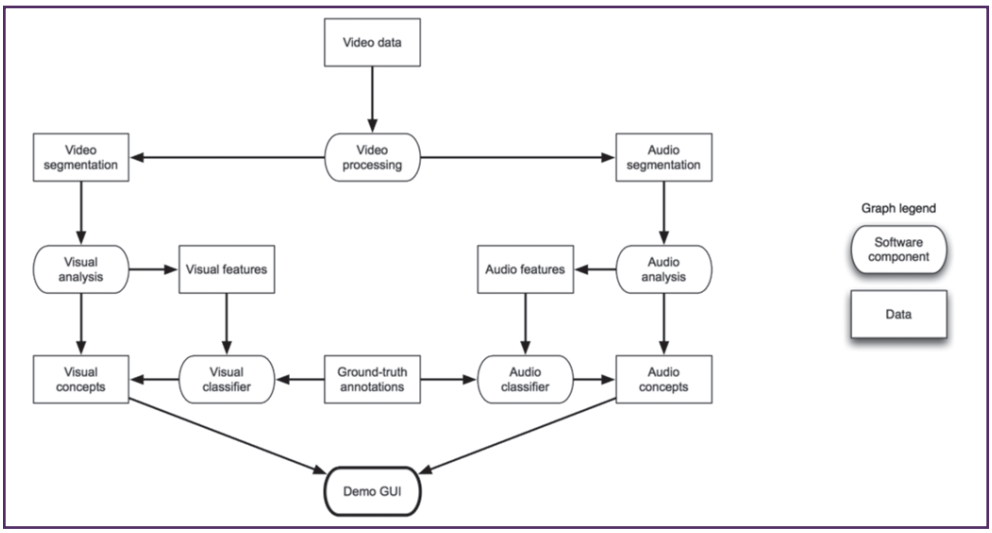

Figure 1. Overview of the system.

The search engine has been evaluated on news broadcast search, video surveillance data, and cultural heritage documentaries repositories. Field trials performed by professional archivists and media professionals have been conducted to evaluate the usability of the system.

The automatic annotation engine

The first step for the automatic annotation engine of VIDI-Video is the video shot segmentation. In this process the video shots are automatically detected, in order to extract key-frames that are then processed to extract the visual features used for concept detection. The shots are then automatically regrouped to create scenes that can be used for video skimming and summarization. The overall process is shown Figure 2.

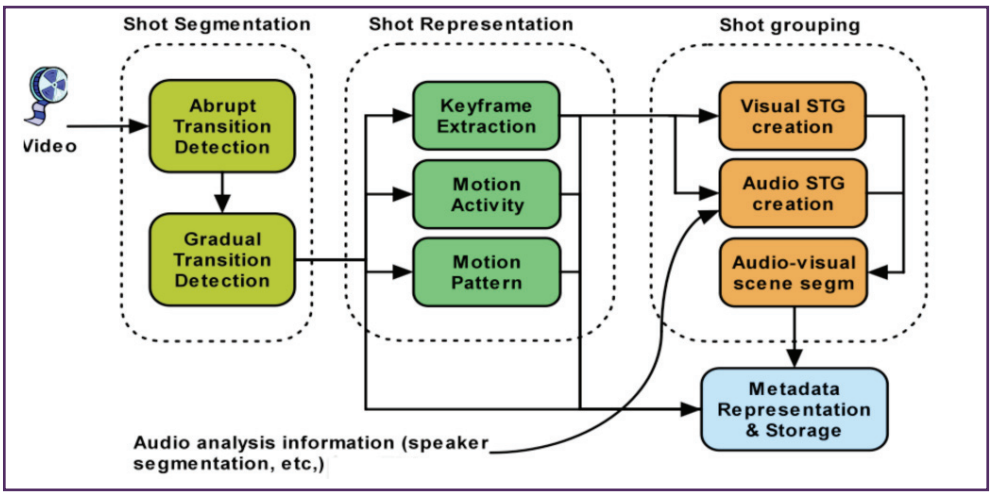

Figure 2. Video shot and scene detection process.

While the detection of abrupt transition, such as *cuts,* is relatively easy, the detection of the gradual transitions (*fades*, *cross dissolves*, etc.) is more complex and may be missed. Within VIDI-Video, a new gradual transition detection algorithm has been developed that uses novel individual criteria that exhibit less sensitivity to local or global motion. The algorithm combines criteria related to changes of color coherence and histograms, as well as luminance (and their multi-scale extensions) using a machine learning technique.

Apart from the advantage of a significantly improved performance, with respect to state-of-the-art, the method does not need any threshold selection, and thus can be applied without requiring users to tune it to their own video collection.

In addition, the creation of video summaries has been improved, fusing both audio and visual information.

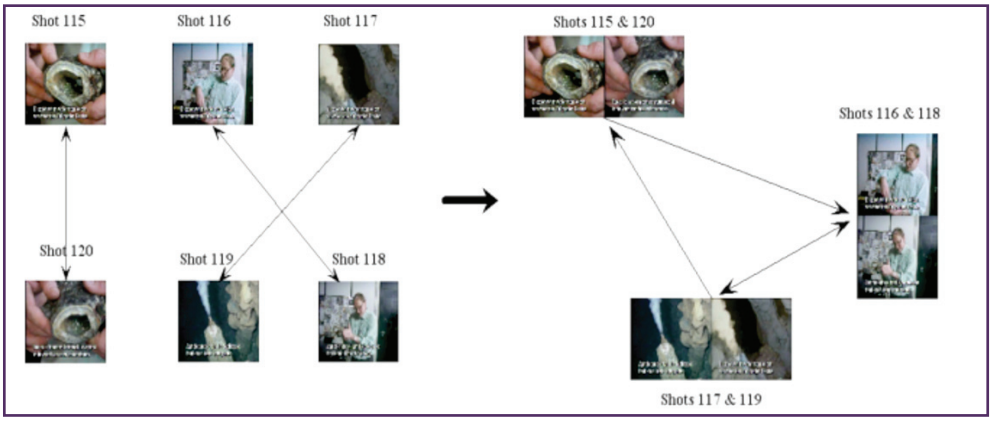

Figure 3. Example of creation of a video summary

The audio analysis component of VIDI-Video (Figure 4) is particularly sophisticated. Apart from the more "traditional" analysis — like automatic speech recognition (implemented for the English language), language identification (the project dealt with videos in different languages, so that it was useful to be able to classify Italian speech from Dutch speech) and topic classification (i.e. understanding the broad argument of a spoken discourse) — the project features a large number of audio events detectors. These detectors recognize audio events related to animals, human activities and tools. Analysis of the presence of human voices is used to infer the presence of dialogues, monologues, etc.

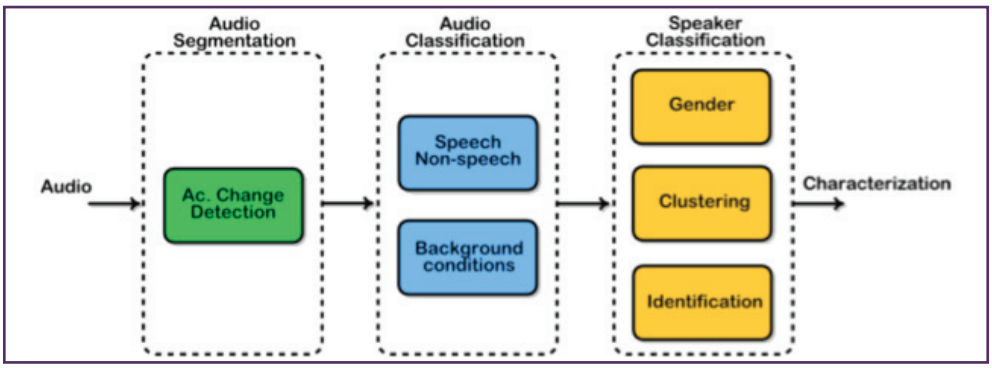

Figure 4. Audio analysis in VIDI-Video

Regarding machine learning techniques, the project has dealt with the development of techniques for learning elements of the thesaurus by combining multimodal features.

Among the main problems that had to be solved we can cite the different modalities and similarity measures, the high computation time and the fact that the training data had highly unbalanced concepts, where some concepts were very rare and others very common.

The solutions developed in VIDI-Video features methods for the fusion of different modalities and information channels, such as audio and visual data, and the development of methods to learn concepts represented by highly unbalanced data sets. Finally, the system is able to learn a large dictionary, such as the 1,000 detector thesaurus, that was the goal of the project. This latter characteristic is extremely important: learning all the independent binary classifiers (i.e. the tools used to recognize if an audiovisual concept is present in a video shot) may become computationally expensive.

At present the VIDI-Video tools are between 5 and 20 times faster than state-of-the-art methods. This has allowed for the increase of the training of the automatic concept detectors, obtaining the best performance in the International competition for visual concept detection PASCAL VOC. The VIDI-Video project has also reached the highest ranking in the TRECVid video retrieval international competition in 2009, thanks to the state-of-the-art visual analysis techniques. A combination of visual content descriptors have been developed, that accounts for both color and luminance information, and that are robust with respect to video framing, lighting, various image distortions, scaling, etc. (Figure 5).

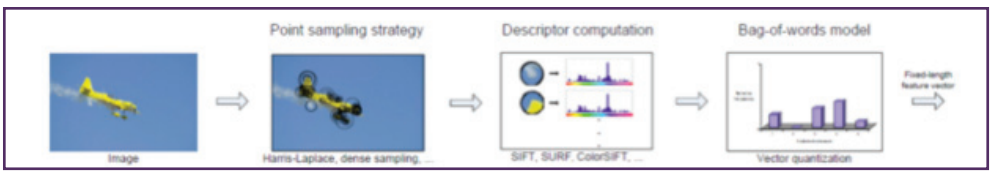

Figure 5. Visual feature extraction

Since this processing is computationally very expensive, a thorough analysis of the bottlenecks has been conducted and a GPU implementation that speeds the whole process by 17 times has been created.

The search engine

Two search engines that aim both at technical and non-technical users have been developed. One has been created as a standalone application, and has been designed for end users that require a fast retrieval process, as it is typical in the media asset management process of a news broadcaster.

The interface allows different query options:

i. query by free text,
ii. query by selecting predicates from a list (according to the annotations stored in the system).

The system also has different visual presentations of query results (Figure 6).

Figure 6. Screenshot of the standalone search engine

The second interface is a web video search engine that allows semantic retrieval by content for different domains (broadcast news, surveillance, cultural heritage documentaries) with query interaction and visualization.

The system permits different query modalities (free text, natural language, graphical composition of concepts using Boolean and temporal relations and query by visual example) and visualizations, resulting in an advanced tool for retrieval and exploration of video archives for both technical and non-technical users. In addition the use of ontologies permits the exploitation of semantic relations between concepts through reasoning. Finally, this web system, using the Rich Internet Application paradigm (RIA), does not require any installation and provides a responsive user interface.

This system is composed of three different interfaces: a GUI to build composite queries that may include Boolean/temporal operators and visual examples; a natural language interface for simpler queries with Boolean/temporal operators; and a free-text interface for Google-like searches. In all the interfaces it is possible to extend queries adding synonyms and concept specializations through ontology reasoning and the use of WordNet. Let's consider, for instance, a query "Find shots with animal": the concept specializations expansion through ontology structure permits the retrieval not only of shots annotated with *animal*, but also those annotated with its specializations (*dogs*, *cats*, *etc.*). In particular, WordNet query expansion, using synonyms, is required when using natural language and free-text queries, since it is not possible to force the user to formulate a query selecting terms from a lexicon, as it is done using the GUI interface.

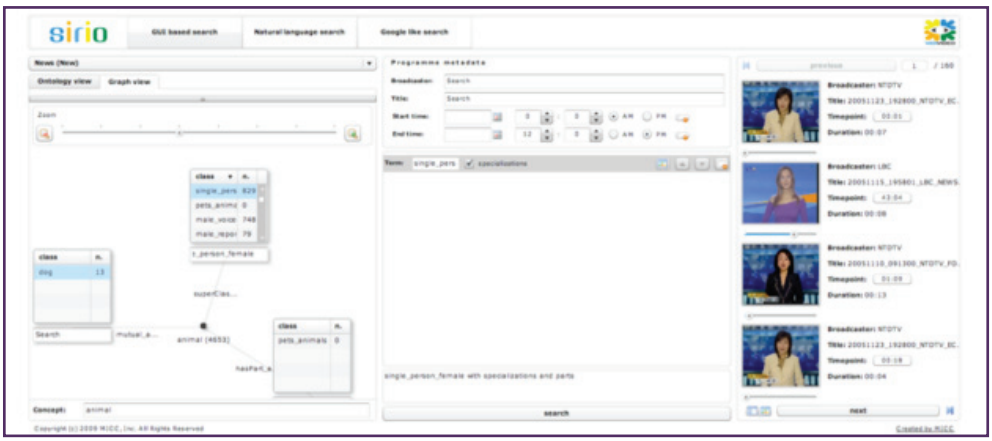

Figure 7. Screenshot of the search engine with a view of part of the ontology

The search engine uses an ontology that has been created automatically from a flat lexicon, using WordNet to create concept relations (*is_a*, *is_part_of* and *has_part*). Concepts, concepts relations, video annotations and visual concept prototypes are defined using the standard Web Ontology Language (OWL) so that the ontology can be easily reused and shared. Queries created in each interface are translated by the search engine into SPARQL, the W3C standard ontology query language.

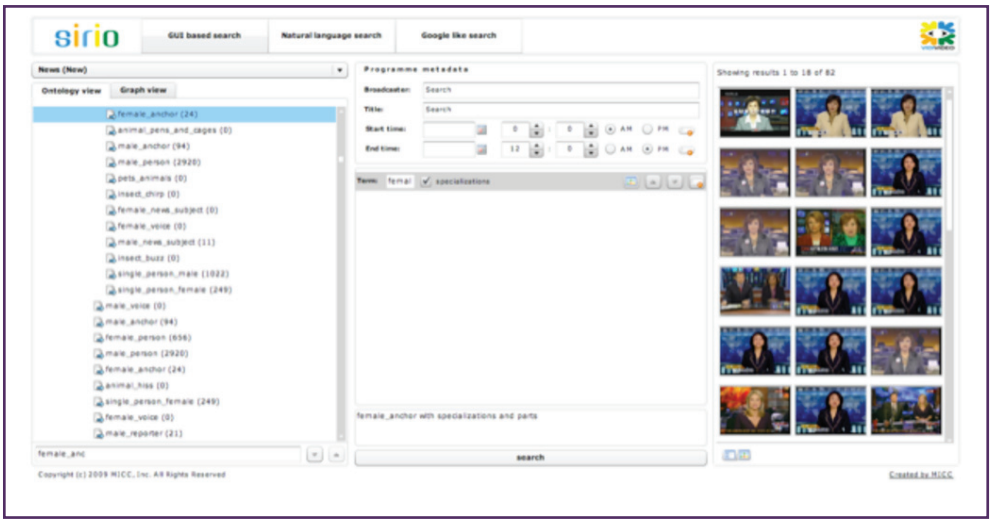

Figure 8. Screenshot of the search engine with a set of results related to the query "female anchor"

The system is based on the Rich Internet Application paradigm, using a client side Flash virtual machine, which can execute instructions on the client computer. RIAs can avoid the usual slow and synchronous loop for user interactions, typical of web-based environments that use only HTML widgets available to standard browsers. This allows for the implementation of a visual querying mechanism that exhibits a look and feel similar to the one of a desktop environment, with the fast response that is expected by users. With this solution, application installation is not required, since the system is updated on the server and can run anywhere regardless of what operating system is used.

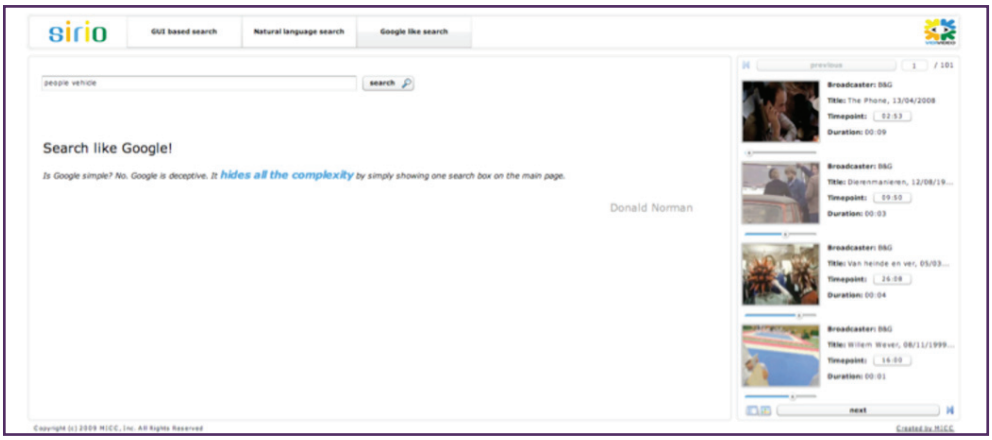

Figure 9. Screenshot of the Google-like query system: just a textbox and a button

The system backend is currently based on open source tools (i.e. Apache Tomcat and Red 5 video streaming server) or freely available commercial tools (Adobe Media Server has a free developer edition). The RTMP video streaming protocol is used.

The search engine has been developed in Java, and supports multiple ontologies and ontology reasoning services. The search results are in RSS 2.0 XML format with paging, so that they can be treated as RSS feeds.

Results of the query are shown in the interface, and the first frame of each video clip of the result set is shown. These frames are obtained from the video streaming server, and are shown within a small video player. Users can then play the video sequence and, if interested, zoom in on each result, displaying it in a larger player that provides more detail on the video metadata and allows better video browsing.

The user interface is written in Adobe Flex and Action Script 3. The GUI interface allows for the building of composite queries that also take metadata into account, as required by professional archivists; the natural language interface allows simple queries to be built with Boolean and temporal relations between concepts; the free-text interface provides the popular Google-like search.

All the modules of the system are connected using HTTP POST, XML and SOAP web services.

The web-based search engine has been tested in a series of field trials conducted by a group of 14 people coming from the broadcasting and media industry and cultural heritage institutions, in Italy (FRD) and in the Netherlands (B&G). The evaluation tests have been carried out in reference to ISO 9241, an industry standard guide on usability.

The overall experience was very positive and the system proved to be easy to use, despite the objective difficulty of interacting with a complex system, for which the users received only a very limited training.

Further information

Within the development of the VIDI-Video project the members of the consortium have produced a large number of scientific papers that describe the main advancements that have been developed. The consortium has also prepared a showcase demo, with videos showing the main functionalities of the system. All this information, as well as contact information, is available on the project web site: http://www.vidivideo.eu.