VIDI-Video Annual Report 2009

www.vidivideo.eu

Video plays a key role in information distribution and access, and it is a natural form of communication for the Internet and mobile devices. The massive increase in digital audio-visual information poses high demands on advanced storage and search engines for both consumers and professional users.

Video search engines are the result of progress in many technologies: visual and audio analysis, machine learning techniques, visualization and interaction. At present, state-of-the-art commercial systems allow for retrieval using keywords found in surrounding text or the speech signal. Only recently they started to allow for retrieval by a small set of semantic concepts such as the presence of a face. VIDI-Video aims to bring semantic access with much more concepts.

To that end VIDI-Video project has realized a sophisticated set of software tools for video annotation and retrieval, that will have a positive impact on cataloging and search practices currently employed in the broadcasting and cultural heritage domain. There will be also an impact in surveillance domain, in which the project has developed a pilot application.

Summary

The VIDI-Video project takes on the challenge of creating a substantially enhanced semantic access to video, implemented in a search engine.
The outcome of the project is an audio-visual search engine, composed of two parts: an automatic annotation part, that runs off-line, where detectors for more than 1000 semantic concepts are collected in a thesaurus to process and automatically annotate the video, and an interactive part that provides a video search engine for both technical and non-technical users.

The automatic annotation part of the system performs audio and video segmentation, speech recognition, speaker clustering and semantic concept detection.
This off-line annotation part has been implemented in C++, and takes advantage of the low-cost processing power provided by GPUs on consumer graphics cards.

The interactive part provides two user interfaces: a desktop-based system and a web-based search engine. 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 different application fields. In addition the use of ontologies (instead of simple keywords) permits to exploit semantic relations between concepts through reasoning, extending the user queries.

The web-based interactive system is based on the Rich Internet Application paradigm, using a client side Flash virtual machine. RIAs can avoid the usual slow and synchronous loop for user interactions. This allows to implement a visual querying mechanism that exhibits a look and feel approaching that of a desktop environment, with the fast response that is expected by users. The search results are in RSS 2.0 XML format, while videos are streamed using the RTMP protocol.

The VIDI-Video system has achieved the highest performance in the most important international contests in object and concept recognition namely PASCAL VOC and TRECVID.

Main activities development

The main activity areas of the project during the third year, each managed by the different partners, are:

- Back-end consolidated software development
- Video processing and creation of the thesaurus
- Development of the user interface prototypes
- Dissemination of the project results

Back-end consolidated software development

After a functional first prototype was implemented and tested, during the second year of the project, a second version -more efficient and with more functionalities- has now been developed.

This back-end automatic annotation software is a consolidation of the different audio and video segmentators, descriptors and detectors developed by the different partners, which also allows the addition of other detectors, such as an OCR system. The different modules are interconnected via an MPEG-7 compliant format for the input and output metadata and the whole system runs on an Ubuntu 8.04 32-bit Linux system.

The software separates completely data storage (both the videos and the metadata extracted from them) so it may be stored on a small, portable space, such as an Ubuntu virtual machine, and work on an external source of data such as a storage server or an external unit.

It runs basically in 2 phases: one for the visual learning tools to train the classifiers for the visual concepts (learning phase, see Fig. 1) from manual annotations and another one for the automatic annotation itself, the results of which are stored into a database (execution phase, see Fig. 2) which will be read by the GUI and delivered to the final user.

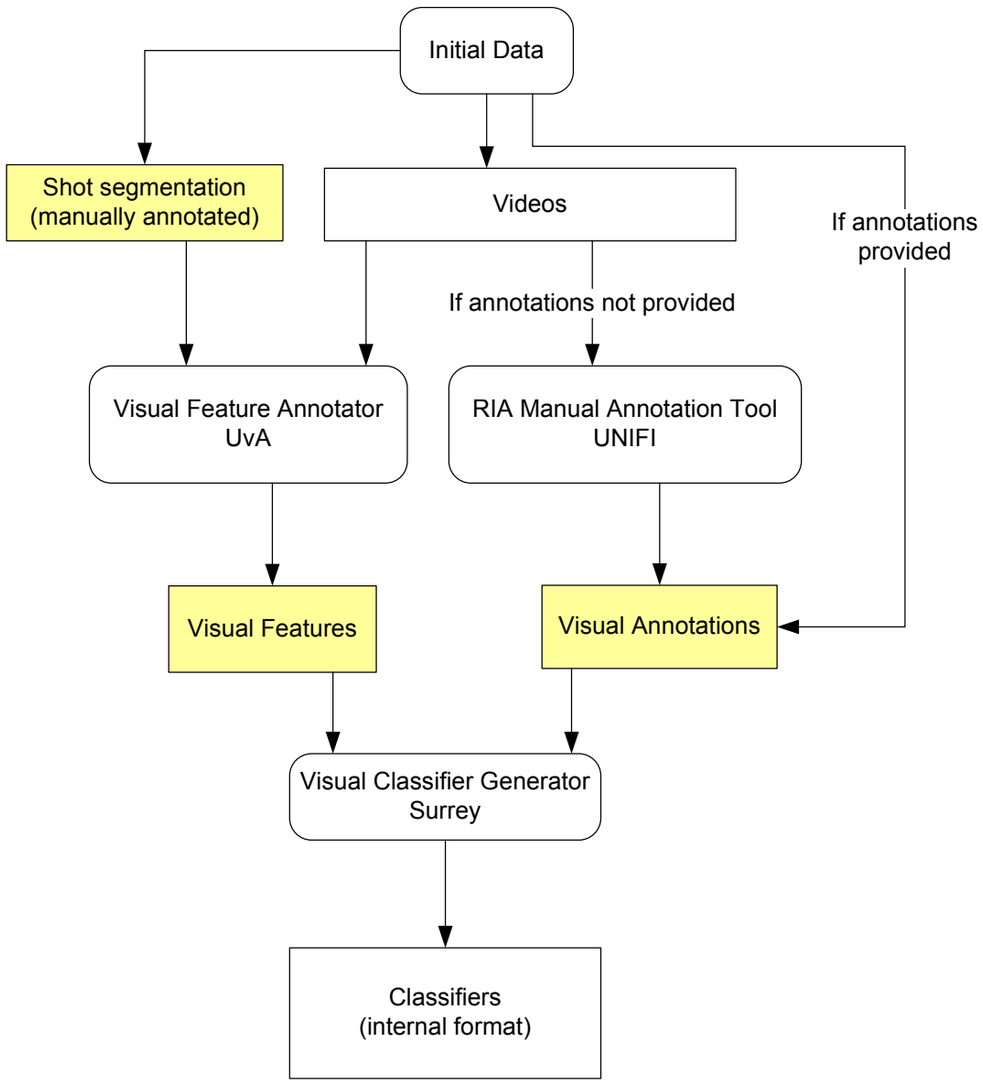

Fig. 1 - Learning phase

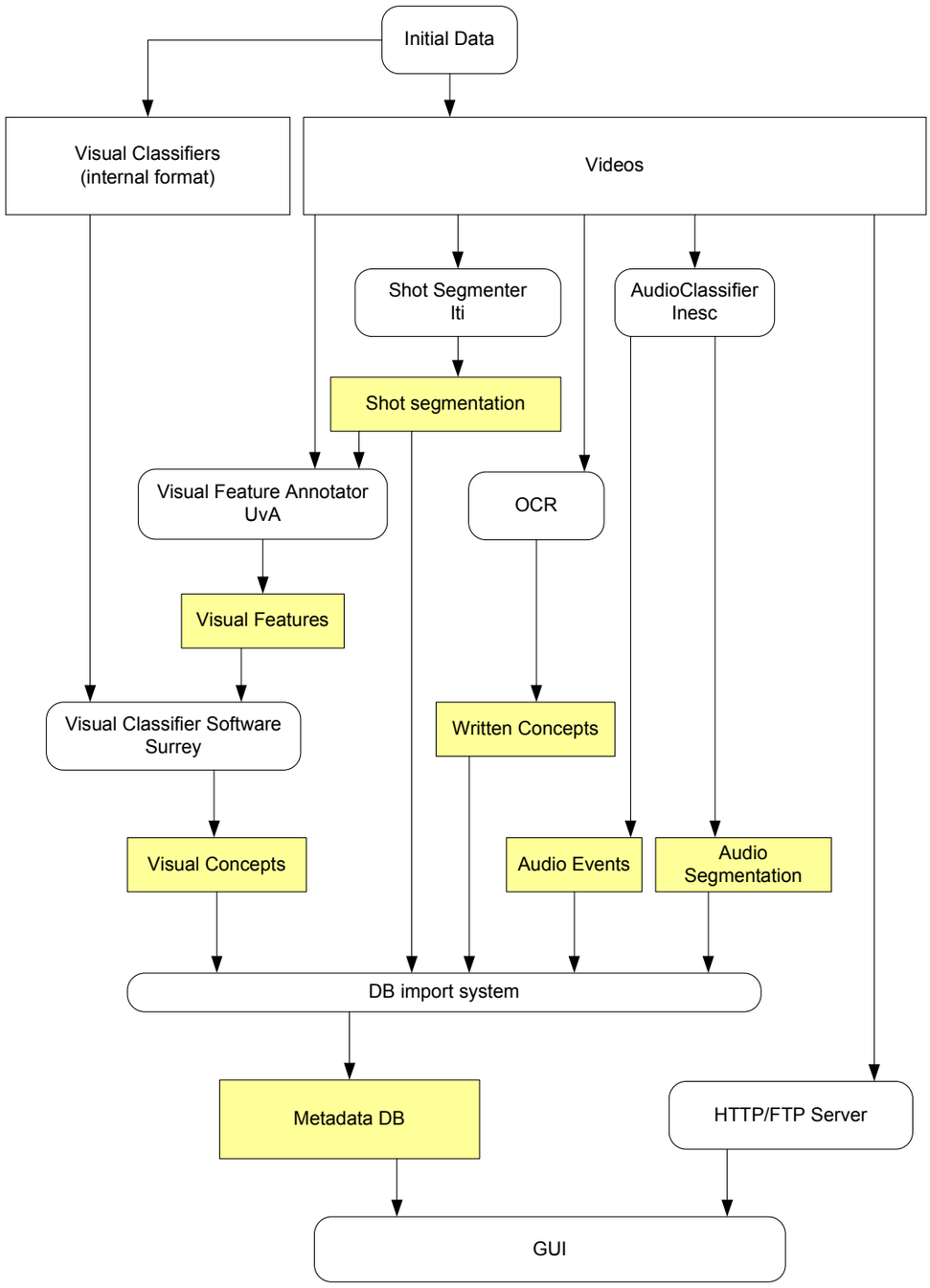

Fig. 2 - Execution phase

Video processing and creation of the thesaurus

A multimodal audio-video segmentation framework, for the processing of video streams, has been studied. The audio part of the thesaurus now has better classifiers for speaker gender recognition and speech/music classification, and new detectors for ambient and sports-related sounds have been added.

An ontology of audio related concepts has been developed, currently including 105 concepts: 75 basic concepts (e.g. birds, horses, buses), 6 aggregated concepts (e.g. animals), 4 music-related concepts (e.g. vocal music), 6 concepts derived from gender/background classification (e.g. child voice), 10 concepts derived from speaker clustering (e.g. dialogue), 2 concepts derived from language identification (e.g. English), 1 concept related to telephone bandwidth audio detection, and 1 concept derived from the transcription (e.g. count down). Classifiers for 62 of these concepts have been properly trained and included in the consolidated software (Fig. 3).

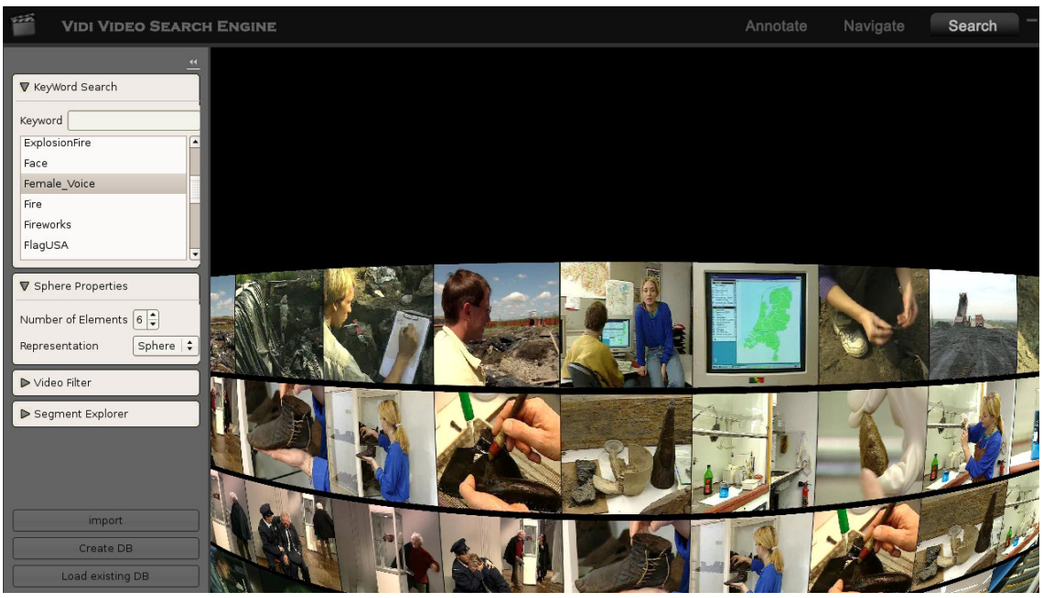

Fig. 3 – Detection results for the audio concept "Female Voice"

For visual analysis more research has been done on spatio-temporal interest points, with the aim to represent human activity events. Possible methods that could be used for classifying events based on such a representation have been investigated.

Regarding the Bag-of-Words approach, which is currently the state-of-the-art method for the classification of visual concepts and as such it is used within VIDI-Video, an effort has been spent on its improvement in terms of computational efficiency and recognition performance.

Regarding efficiency, a factor 10 improvement in speed, with no loss in accuracy, or a factor 30 speed improvement with a 17% accuracy loss, is achieved. Further improvements have been obtained by a careful analysis of the processing bottlenecks, and implementing parts of the algorithms to exploit the general-purpose GPU parallelism, so to obtain 20-45x speed improvements for the time-consuming steps in the overall process.

Regarding the goal of improved recognition performance a theoretical work has

investigated the relative importance of object and context patches within the Bag-of-Words framework. It was concluded that being able to separate context from object patches would improve retrieval rates by 26%. Other work has dealt with the improvements of robust color features that can be used recognition and retrieval of images and videos.

The system developed has been tested within important international benchmarks. Regarding TRECVid 2009, the concept detectors have been run on over 1 million frames in the dataset (result: best overall run), with the highest score for 10 out of 20 concepts. Regarding PASCAL VOC 2009, the new color filters have been applied for the classification task (result: named one of the winners of the task). Further, we also participated in the large-scale visual concept detection task (LS-VCDT) of ImageCLEF 2009 (result: best overall run), obtaining the highest performance for 40 out of 53 concepts.

Development of the user interface prototypes

The work on the three user interfaces developed in the previous year has continued. The web-based interface (Fig. 4 and Fig. 5) 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.

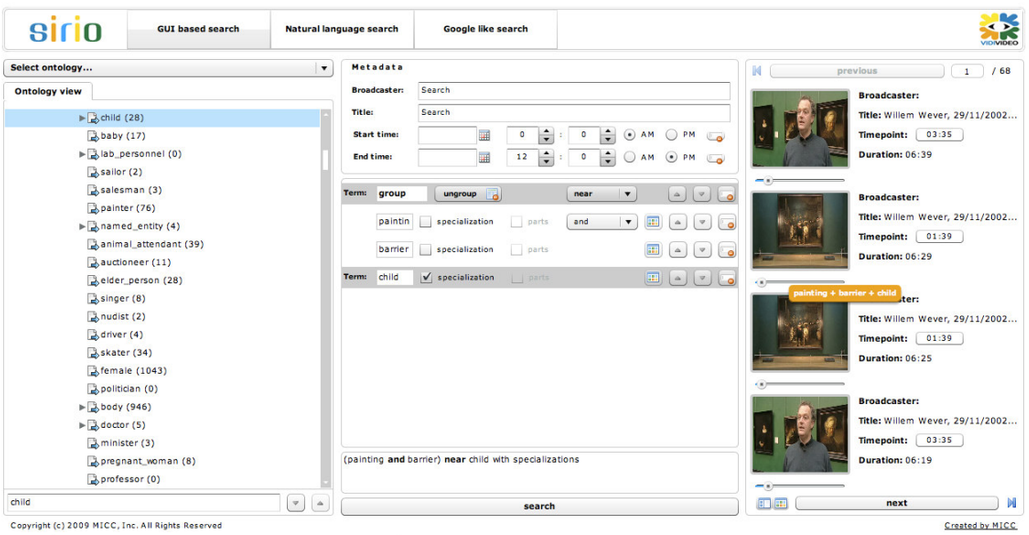

Fig. 4 - composite query with drag&drop

In all the interfaces it is possible to extend queries adding synonyms and concept specializations through ontology reasoning and the use of WordNet. Consider, for instance, a query “Find shots with animals”: the concept specializations expansion through ontology structure permits to retrieve not only the 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 is done using the GUI interface.

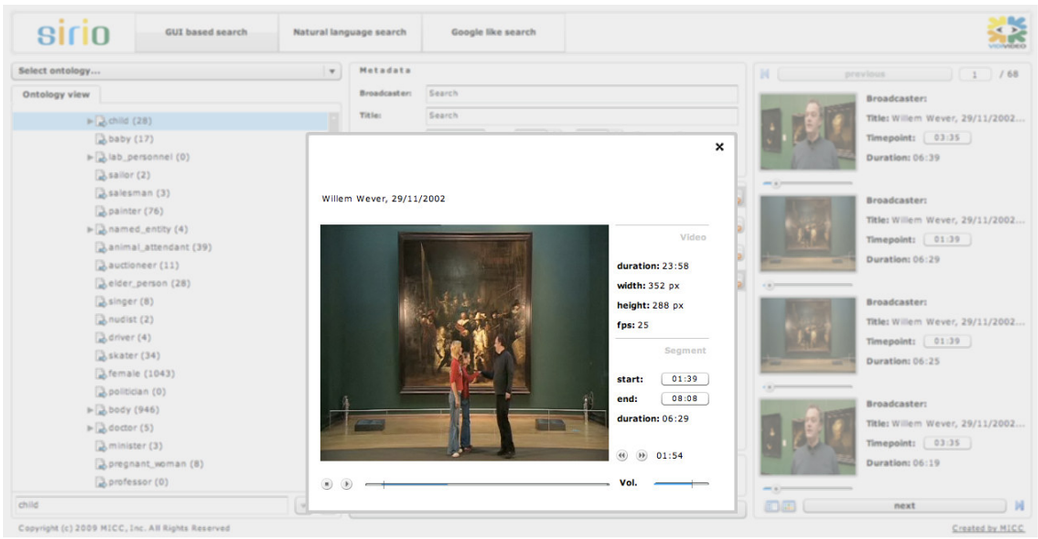

Fig. 5 - streaming video player, to inspect the results of the query

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 the HTML widgets available to standard browsers.

Using RIAs allows implementation of a visual querying mechanism exhibiting a look and feel approaching that of a desktop environment, with the fast response that is expected by users.

Dissemination of the project results

VIDI-Video has distinguished three main user groups that could benefit from the outcomes of the project: Broadcast Archives, Cultural Heritage Institutions and Video-surveillance domain.

Along with these target groups of potential users there is also the scientific community, in particular represented by other IST projects, networks of excellence, and other RTD projects, and the private sector represented by IT companies, media industry and other players in the AV market.

For each of the above-mentioned target groups, in the last year, different VIDI-Video partners have organized or participated in specific events/conferences to raise awareness and disseminate the project results.

In particular, for the broadcast and audiovisual archives and cultural heritage professionals the VIDI-Video project was presented during the **40th IASA** (International Association of Sound and Audiovisual Archives) **annual conference** in Athens, in September 2009.
(http://www.iasa2009.com)

For the video-surveillance community, the project was presented during an Italian national **workshop on video-surveillance for urban safety**, held in Modena on February 2009. More than 300 participants, mainly local government officials, police officers, and lawyers attended to the workshop and to the demo of the VIDI-Video prototype for video-surveillance.
(http://imagelab.ing.unimore.it/videosorveglianza2009)

On May 2009, the VIDI-Video project was presented during the **CHORUS Final Conference** in Brussels. CHORUS is a European Coordination Action, which aims at creating the conditions of mutual information and cross-fertilization between the European projects dealing with Multimedia Content Search Engines.
(http://www.ist-chorus.org/conference.asp)

The VIDI-Video web based prototype systems have been presented also during the **ACM Multimedia 2009** conference in Beijing in October.
(http://www.acmmm09.org)

The final showcase of the project will be held during the **4th International Conference on Semantic and Digital Media Technologies** (SAMT '09) in December in Graz. The conference targets at narrowing the large disparity between the low-level descriptors that can be computed automatically from multimedia content and the richness and subjectivity of semantics in user queries and human interpretations of audiovisual media: the Semantic Gap. The VIDI-Video consortium is sponsoring the conference, and all the project results and achievements will be presented during a dedicated session on the last day of the conference.
(http://www.samt2009.org)

Further information about these and other project activities can be found at http://www.vidivideo.eu