VIDI-Video Annual Report 2010

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 has brought 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 the surveillance domain, in which the project has developed a pilot application.

Summary

The VIDI-Video has taken 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 overall structure of both parts are depicted in figure 1 and 2.

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 (see **Fig.** 3.) and a web-based search engine (see Fig. 4 and Fig. 5.). 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 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.

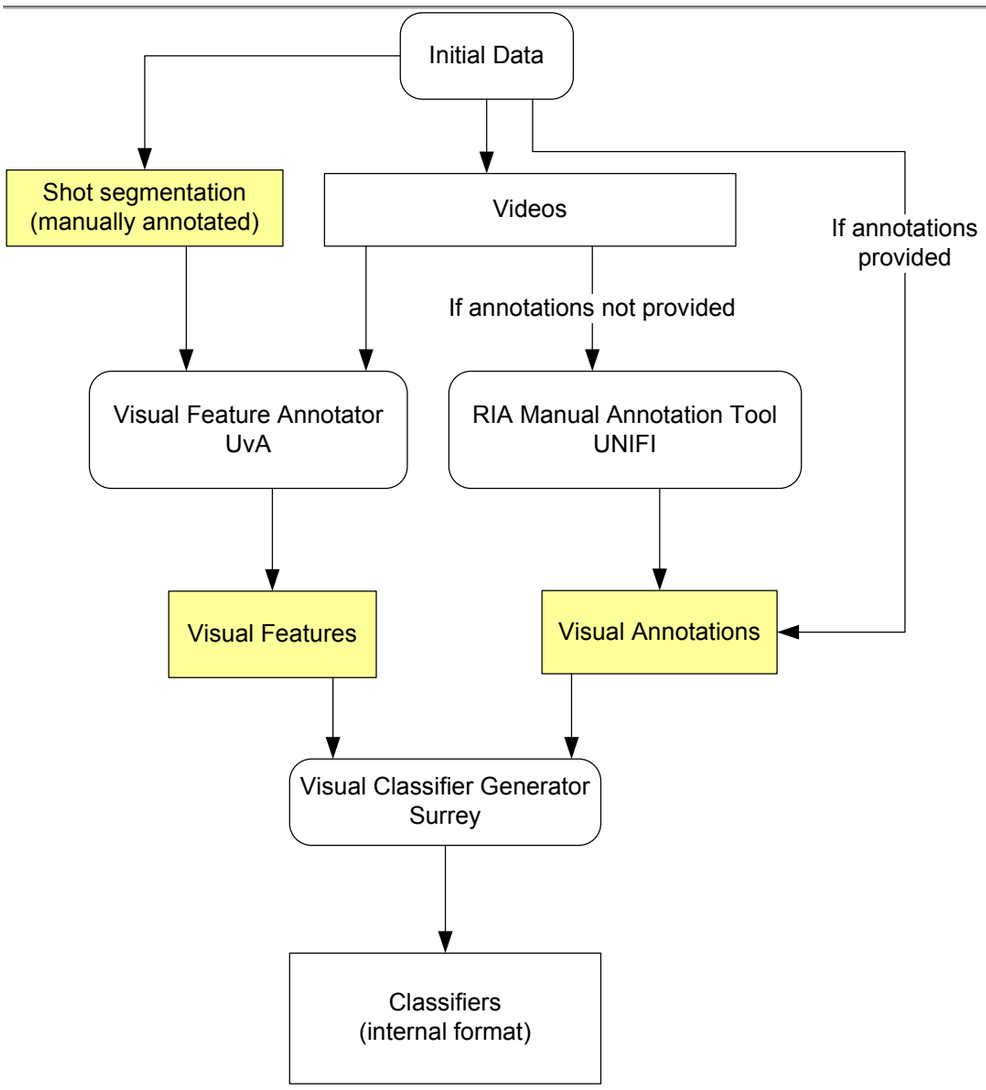

Fig. 1 - Learning phase

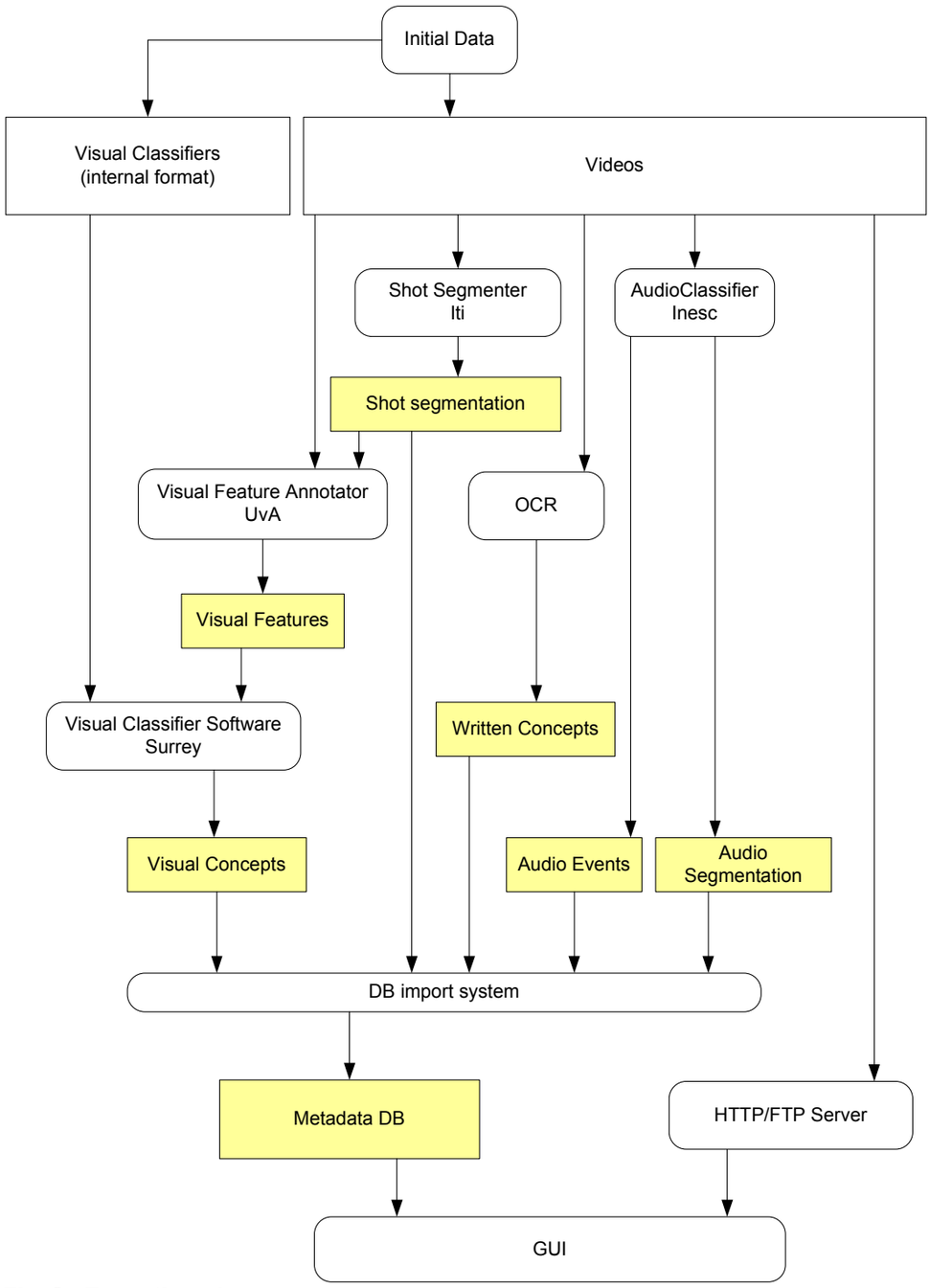

Fig. 2 - Execution phase

The user interface prototypes

The stand alone interfaces is depicted in Fig.3. where the web-based interface are presented in Fig. 4 and Fig. 5. The web interface 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.

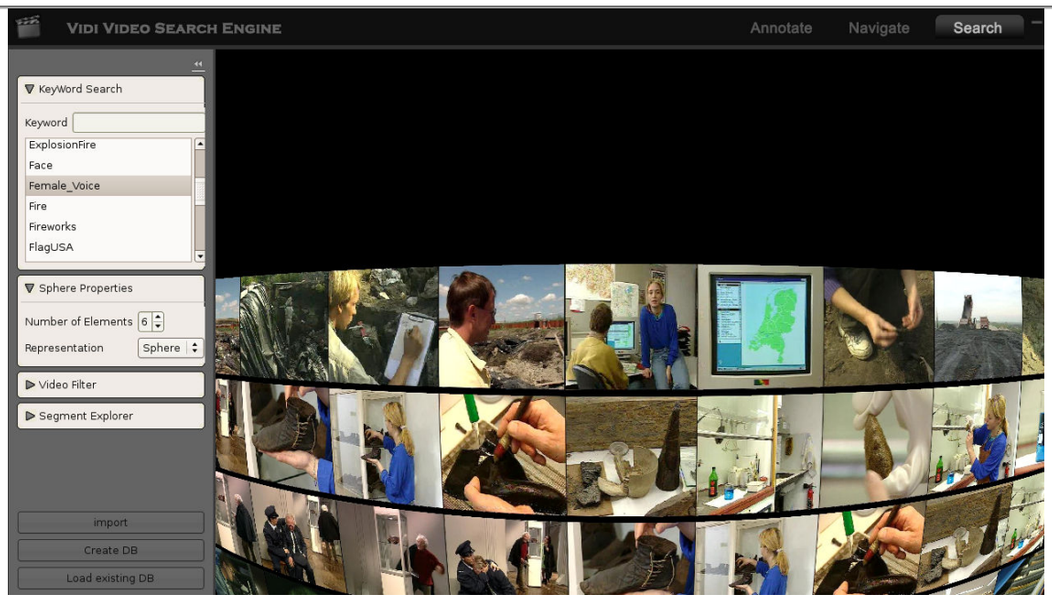

Fig. 3 – Detection results for the audio concept "Female Voice"

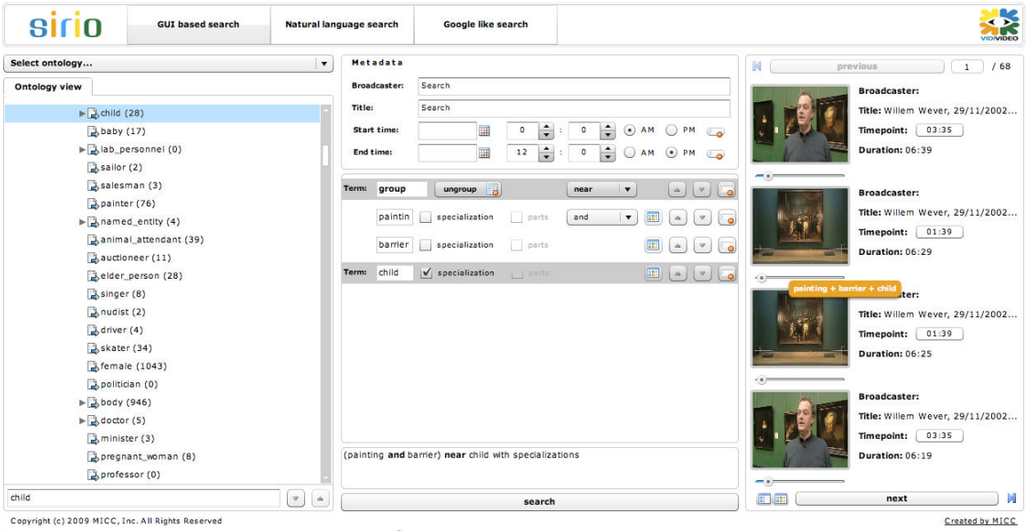

Fig. 4 - composite query with drag&drop

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.

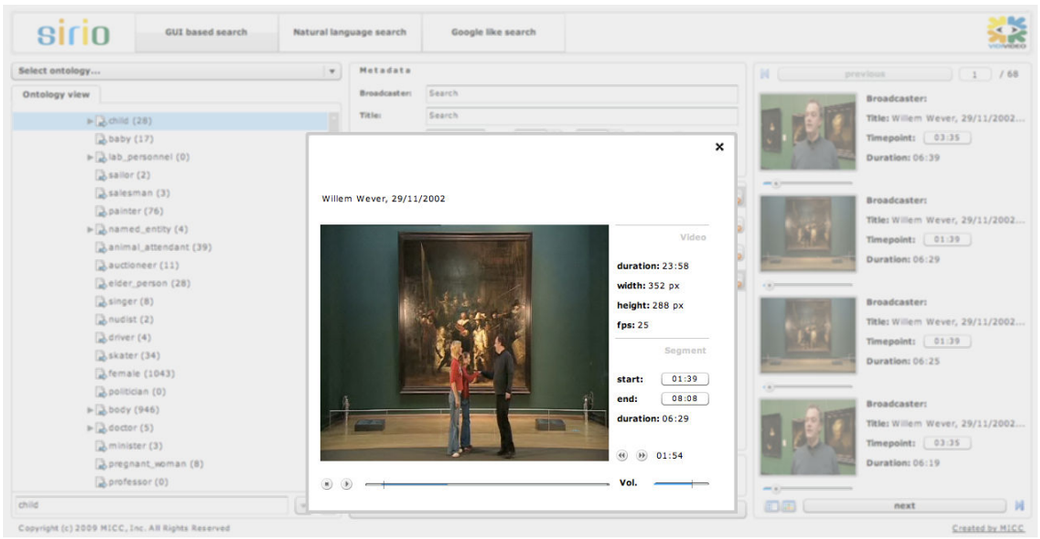

Fig. 5 - streaming video player, to inspect the results of the query

Dissemination of the project results

The final showcase of the project has been the **4th International Conference on Semantic and Digital Media Technologies** (SAMT '09) in December in Graz. The conference targeted 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. All the project results and achievements have been 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