VIDI-Video: Interactive semantic video search with a large thesaurus of machine-learned audio-visual concepts

Marco Rendina , Fondazione Rinascimento Digitale, Marco Bertini, Università di Firenze, Italy

Video is vital to society and economy. It plays a key role in information distribution and access and it will be also 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 the consumer and the professional will need advanced storage and search technology for the management of large-scale video assets.

Current search engines, however, mostly rely on keyword-based access leaving semantic access to the data to research. VIDI-Video project takes on the challenge of creating a substantially enhanced semantic access to video. The project aims to integrate and develop state of the art components from machine learning, audio event detection, video processing, interaction and visualization into a fully implemented audiovisual search engine combining large number of categories and exploiting the interclass similarities as well as using the information from different sources: metadata, keyword annotations, audiovisual data, speech, and explicit knowledge.

VIDI-Video aims at boosting the performance of audiovisual search by forming a 1,000 detectors thesaurus aiming to localize the corresponding semantic concepts in the audio, visual or combined stream 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 audiovisual content. In order to reach this goal of semantic understanding, VIDI-Video will improve on machine learning techniques, visual and audio analysis techniques and interactive search methods. The approach is to let the system learn many, mostly weak, semantic detectors instead of modeling a few of them carefully. These detectors will describe different aspects of the video content. In combination they will render a rich basis for interactive access to the video library. Concrete outputs will be an audiovisual search engine, consisting of a learning part and a runtime system. The learning part will consist of units for video processing, visual analysis, audio analysis, and learning integrated feature detectors. The runtime system includes the same video, audio and visual units and the thesaurus of semantic concept detectors updated in performance after each round of training. The system will have also an ontology-based web user interface based on the R.I.A. paradigm, and also a stand-alone interface. The search engine will be evaluated on news broadcast search, video surveillance data, and cultural heritage documentary repository. Mobile video applications can be also part of a future scenario of the current technology. The consortium presents excellent expertise and resources: the *machine learning* with active 1- class classifiers, to minimize the need for annotated examples, is lead by the University of Surrey (UK). *Video stream processing* is lead by CERTH (Greece). Another component is *audio event detection,* lead by INESC-ID (Portugal). *Visual image processing* is lead by the University of Amsterdam (Netherlands). The University of Florence (Italy), leads the efforts in *interaction*, and CVC (Spain) leads *software consolidation*. Finally, Beeld & Geluid (Netherlands), and FRD (Italy), as application stakeholders, provide *data* and perform *evaluation* and *dissemination*.