OLIVE: speech based video retrieval

Klaus Netter
Language Technology Lab
German Research Center for Artificial Intelligence – DFKI GmbH
Stuhlsatzenhausweg 3, D-66123
Saarbrücken, Germany.
E-mail: netter@dfki.de

Franciska de Jong
TNO/University of Twente
address: Department of Computer Science
P.O. Box 217, 7500 AE Enschede
The Netherlands.
E-mail: fdejong@cs.utwente.nl

ABSTRACT

This paper describes the Olive project which aims to contribute to the needs of video archives by supporting the automated indexing of video material on the basis of human language processing. Olive develops speech recognition to automatically derive transcripts for the sound track, thus generating time coded linguistic elements which are the basis for text-based retrieval functionality.

Keywords: language technology, content-based video retrieval, speech recognition

1 INTRODUCTION

In archives detailed documentation and profiling of the archived material is a prerequisite for an efficient and precise access to the data. While in the domain of textual digital libraries advanced methods of information retrieval can support such processes, there are so far no effective methods for automatically profiling, indexing, and retrieving image and video material on the basis of a direct analysis of its visual content. Although there have been some advances in the automatic recognition of images, these are still so limited that they will not provide a sufficiently robust basis for effectively profiling large amounts of visual data. Instead Olive uses natural language as the media interlingua, focusing on technology for processing the sound track. It is a follow-up of the Pop-Eye project (http://pop-eye.tros.com/) which takes subtitles as starting point.

2 USER NEEDS

The primary users of Olive are two broadcast organisations (ARTE and TROS), as a national audio-video archive (INA) and a large service provider for broadcasting and TV productions (NOB). For all of these institutions archiving of video productions plays an important role, be it for the purpose of re-broadcasting or reselling existing productions, for reusing part of the material in new productions or for generally supporting research in video data bases. In particular, the latter two functions make it very important that the customers of the archives have maximally detailed access to the content of the video material.

Reusing parts of existing material can reduce the production costs considerably and therefore makes it highly desirable that the full and detailed content of a video be documented and accessible without having to view the entire video. This implies that indexes to video's would have to disclose not just the video production as a whole, but also fragments of the material via their timecodes. As generating the necessary content descriptions for large numbers of video shots per production is very costly and labor-intensive, automated indexing is a way to meet the demands of present day multimedia archives.

Olive aims to develop a system which automatically produces indexes from a transcription of the sound track of a programme. In addition the Olive system will provide access to the digitised video material through some intranet or even the internet. As a result user should be able to query a digital video library, browse through the returned descriptions and then download and pre-view the relevant sequences.

3 BASELINE TECHNOLOGY

To answer such problems and demands as just described Olive attempts to provide online access to video material on the basis of linguistic material associated with the visual data. The linguistic data associated with a video basically come in two classes. They are either linked to the video time code or not. Among the former are subtitles and of course the spoken word itself. In addition to disclosure technology for the tasks to be performed by any retrieval system, Olive will develop *speech recognition* for German and French for the automatic generation of timecoded transcriptions of the sound track. Non-time coded texts will be timecoded with *alignment techniques*. In addition Olive will also apply *translation technology*.

3.1 SPEECH RECOGNITION & ALIGNMENT

Currently, speech technology is still somewhat limited and does not guarantee completely domain- and speaker-independent reliable recognition. However, it has to be kept in mind, that for the purpose of indexing and retrieval a 100% recognition rate is not absolutely necessary, since not every word will have to make it into the index, and not every expression in the index is likely to be queried. In addition, speech recognition can also be used as a secondary means to support automatic time coding of the second class of data, as for example manual transcriptions. The cleaner and more reliable transcriptions can be used as the basis for indexing. The necessary time-coding can then be derived by automatically aligning the result of speech recognition with such a transcription. Basically the same method can be used if there are production scripts or other types of descriptions reflecting the time line and the spoken word.

3.2 TRANSLATION

Following the approach developed within Twenty-One (http://twentyone.tpd.tno.nl/) functionality will be added to support cross-language information retrieval. For example, video's with a German soundtrack will be accessible via queries in any of the languages French, English, Dutch and German.

3.3 INHERENT LIMITATIONS

It should be clear, of course, that the discourse and linguistic data associated with a video will not always be a direct reflection of the images and the visual content of the video. In particular, there will be a broad range of variation between more descriptive texts, like documentaries, where the commentary refers to and explains the visual content, and programmes of the drama type, where the dialogue and discourse complements the visual content. Thus, the approach taken in the projects will have some clear limitations, and future experience and evaluation will have to show for what type of programmes the approach is most suitable.

4 PROJECT INFORMATION

The users in the Olive consortium are two television stations, comprising ARTE (Strasbourg, France) and TROS (Hilversum, Netherlands), as well as the French national audio-video archive, INA/Inatheque in Paris, France, and a large service provider for broadcasting and TV productions, viz., NOB in Hilversum, Netherlands.

The system will be implemented through the co-operation of several organisations: TNO-TPD Delft, the project co-ordinator which brings in the core indexing and retrieval functionality, VDA BV Hilversum building the video capturing software, the University of Twente and the LT Lab of DFKI GmbH Saarbrücken, responsible among others for the language technology, the University of Tübingen, carrying out the evaluation in Pop-Eye, CNRS LIMSI and Vecsys SA Paris which are developing and integrating the speech recognition modules, respectively.

Olive (LE4-8364) is funded by the European Commission under the Telematics Application Programme, sector Language Engineering. The project started in 1998 and will last until 2000, More information about Olive can be found under http://twentyone.tpd.tno.nl/olive.