OLIVE: Speech Based Video Retrieval

Franciska de Jong[1], Jean-Luc Gauvain[2], Jurgen den Hartog[3], Klaus Netter[4]

[1] TNO/University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
[2] LIMSI-CNRS, BP 133, 91403 Orsay, France
[3] TNO-TPD, Stieltjesweg 1, 2628 CK Delft, The Netherlands
[4] DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

Abstract. This paper describes the OLIVE project which aims to support automated indexing of video material by use of human language technologies. OLIVE is making use of speech recognition to automatically derive transcriptions of the sound tracks, generating time-coded linguistic elements which serve as the basis for text-based retrieval functionality. The retrieval demonstrator builds on and extends the architecture from the POP-EYE project, a system applying human language technology on subtitles for the disclosure of video fragments.

1 Introduction

In archives of all kinds, detailed documentation and profiling of the archived material is a prerequisite for 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 of course advances in the automatic analysis and recognition of images, these are still so limited that they do not provide a sufficiently robust basis for profiling large amounts of homogeneous visual data. Instead the LE-4 OLIVE project uses natural language as the media interlingua, making the assumption that, currently, the detailed content of video material is best disclosed through its linguistic content. For this purpose OLIVE is focusing on speech technology processing of the sound track, but also taking into account other linguistic material associated with video documents.

The main objective of the OLIVE project is to develop a radio and video archiving and retrieval tool that will facilitate efficient access to large libraries of multimedia material. The project is developing and testing a prototype which automatically partitions the audio channel and transcribes the speech portions producing a time-coded orthographic transcription. From the transcript an index of appropriate terms is derived with each phrase being linked to specific time points of the video programme. OLIVE is developing tools to support users in searching for material via natural language queries, including cross-lingual indexing and access based on offline machine translation of the archived documents, and online query translation.

The OLIVE consortium is comprised of users and technology providers and integrators. The primary users of OLIVE are two broadcast organisations (ARTE and TROS), a national audio-video archive (INA) and a large service provider for broadcasting and TV productions (NOB). Technology providers include TNO (Coordinator), University of Twente and DFKI for natural language processing, LIMSI-CNRS for speech recognition technology, and VECSYS and VDA for integration.

This paper presents an overview of the project goals, both from the perspective of the users and the technology developers. Section 2 addresses the user needs, and Section 3 describes the core human language technologies used for speech recognition, indexation and retrieval. Finally in Section 4 some more detailed project information is given, including an overview of the major achievements thus far in the project and a short description of the demonstrator that has been built.

2 User Needs

The prime interest of the OLIVE users is to obtain an efficient, detailed and direct access to their video archives. The users help to guide the project, specifying the desired system functionalities, and will evaluate and test the prototype in their own working environments.

For the user institutions, disclosure of video material plays an important role, be it for the purpose of re-broadcasting or re-selling existing productions, for re-using part of the material in new productions or for supporting research in video databases. With rising production costs, re-broadcasting is an important means of writing off the costs over time. Re-selling material, in particular across country and language boundaries, is likewise an additional source of income, which makes multilingual access to archives a desirable property. Re-using and integrating existing material can reduce the cost for a new production by a factor of 10 or more. Enabling detailed research is one of the main functions of public audio-video archives, such as INA, but can also play a role for producers and editors in TV stations.

Most of these needs make it very important that the users of the archives have direct access to the content of the video material without having to view the entire document. This implies that indexes to videos have to refer not just to the video production as a whole, but also to fragments of the material via their time code.

When video archives are disclosed, this is typically carried out by archivists and documentalists, who view the video and in parallel note its content through keywords or descriptive expressions.[1] While this method is maximally precise and detailed for the purpose of capturing the visual content of a video, it is also extremely time and cost consuming. For the detailed disclosure of a video, a ratio of 1:15 can be assumed, i.e., for one hour video up to fifteen hours of description time can be necessary. It is quite clear that such a method can only be applied to selected productions, and that the vast majority of material cannot be disclosed on this basis at all.

Olive aims to support such human archiving processes by developing a system which automatically produces full text indexes from a transcription of the sound track of a programme. This indexing method is meant to complement traditional methods by offering another, and in some cases an exclusive information channel into the video material.

In addition to the detailed content disclosure, the Olive system will also provide access to the digitised video material through network technology, specifically web browsing. This will answer the growing demand to preview material remotely, before actually obtaining the material from the archives. Rather than having to collect the material for browsing, a user will be able to query a digital video library from his desktop, browse through the returned descriptions and then download and pre-view the relevant sequences. The overall philosophy in searching for video material is that the user can narrow down his search by abstracted information in the form of index terms, text passages, transcriptions or subtitles, time-abstracted story-boards or sequences of reduced stills, in order to finally focus in on the actual video sequence.

3 Core Technologies

To answer the problems and demands described above, Olive attempts to provide online access to video material on the basis of linguistic material associated with the visual data. The linguistic data connected with a video basically can be divided into those which are inherently linked to the temporal dimension of the video and those which are not. Among the former are subtitles, which carry some invisible time code and of course the spoken word itself which is time-coded through the alignment of the sound track with the video signal.

One of the main technical tasks to be faced by Olive is therefore to segment and process the linguistic data such that each linguistic expression which qualifies as an index term can be directly associated with the time code referring to a corresponding video sequence. This is trivially achieved if the linguistic expression is already in a time-coded textual format, as in the case of subtitles. For all other data, the time-code and the textual representation has to be derived. Speech recognition in French and German, which is being used to automatically generate time-coded transcriptions of the sound track, is therefore one of the core technologies to provide the necessary information.

For non-time-coded texts, such as scripts, manual transcriptions produced for translation or subtitling, a time-coding is being derived by automatic aligment with the time-coded data. Since non-time coded data typically consists of manually produced and controlled textual material, the quality of the index terms from such data could even be more reliable than the one derived from speech transcriptions.

For the retrieval functionality Olive is building on some of the core functions of a search engine which was originally developed in theTwenty-One project(http://twentyone.tpd.tno.nl/). This search engine, which was also used within the Pop-Eye project, is described in more detail below. To support cross-lingual search and retrieval, Olive is applying translation technology both for offline document translation and for online translation of query terms.

[1] Institutions which carry out detailed disclosure processes are for example German ARD or Belgian VRT TV stations.

3.1 Speech Recognition

To address the various user needs, Olive supports different transcription modes: segmentation, guided and fully automatic transcription. For the segmentation task, a perfect transcription of the spoken data is assumed, and this transcript is time-aligned with the acoustic signal. However, existing transcripts are unlikely to be exact transcripts of what was said and/or may only be partial transcripts. In the worst case there may be only accompanying texts, such as a summary. The most efficient means of generating time-aligned transcripts is to use the associated text to guide the search during recognition, which is what can be qualified as informed speech recognition. When time-coding is needed for the original document, the speech recognizer output has to be aligned with the available documents. Confidence scores are associated with each hypothesized word to allow further processing steps to take into account the reliability of the candidates.

The state-of-the-art speech recognizer developed at LIMSI[1] is a continuous mixture density, tied-state cross-word context-dependent HMM system with a 65k word trigram language model. Decoding is carried out in multiple passes, incorporating cluster-based test-set acoustic adaptation.

Prior to word recognition, the acoustic signal is partitioned into homegenous segments, and appropriate labels are associated with the segments[2]. This partitioning algorithm first detects (and rejects) non-speech segments using Gaussian mixture models (GMMs). An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

The speech recognizer[1] uses context-dependent triphone-based phone models, where each phone model is a tied state left-to-right CD-HMMs with Gaussian mixtures and the tied states are obtained by means of a phonemic decision tree. Word recognition is performed in three steps: initial hypothesis generation, word graph generation, and final hypothesis generation. The initial hypotheses are used for cluster-based acoustic model adaptation.

Taking advantage of the corpora available through the LDC, the speech recognizer[1, 2] has been developed and tested on American English. The acoustic models are trained on 150 hours of transcribed audio data, with the language models trained on 200M words broadcast news transcriptions and 400M words of newspaper and newswire texts. Using broadcast data collected in Olive, LIMSI has ported its American English system to French and a port to German is underway.

Experiments with 700 hours of unrestricted broadcast news data indicate that word error rates around 20% are obtained for American English. Preliminary experiments in French and German indicate that the word error rates are somewhat higher, which can be expected as these languages are more highly inflected than English, and less training data are available. 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. Research into the differences between text retrieval and spoken document retrieval indicates that recognition errors do not add new problems for the retrieval task[6].

3.2 Indexing and Retrieval

The retrieval functionality of Olive builds on the technology developed for Twenty-One which was the first on-line search engine in Europe supporting cross-language retrieval (accessible since 1996). The system supports the automatic disclosure of information in a heterogeneous document environment, covering documents of different types and languages.

The Twenty-One retrieval technology was evaluated on two tasks of the international IR evaluation conference TREC-7. Both in the main task and in the cross-language task, the Twenty-One system performed at the level of today's world leading experimental IR systems. Cf. [5].

The objective of the Twenty-One system was to develop domain-independent technology to improve the dissemination level of digitised and non-digitised multimedia information. It has set a baseline for a series of EU-funded projects developing multimedia indexing tools. An application of the system in the domain of sustainable development can be inspected at: http://twentyone.tpd.tno.nl./twentyone/ Cf. also [3] and [4].

The language elements in the documents to be disclosed are the basis for the automatic generation of a text based index that enables the kind of functionality commonly known as full text retrieval. This provides users access to information not just via a controlled set of search terms, but via any word in the document. It allows users not only to look for entire documents, but also for information within the documents.

The retrieval system thus consists of two crucial sets of software: (i) software to disclose multimedia information, including a series of natural language processing modules and (ii) software to retrieve multimedia information (with state-of-the-art browsing applications) from remote or local servers, or from a local CD-ROM. The latter contains a search kernel supporting several query modes and interface languages.

The disclosure subsystem builds on linguistic software which includes morphological analysis and part-of-speech tagging, parsing (noun phrase extraction) and translation. This goes beyond the analysis parts of standard full text retrieval systems, in as far as such systems often do not even comprise lemmatisation let alone phrasal structuring in their analysis part. The parser output consists of a version of the original document in which the noun phrases (NPs) or other phrasal units – which are considered to be potential index terms – have been marked. For the output of the speech recogniser, linguistic analysis and segmentation at a higher (phrasal, clausal or sentential) level is even more important, as here the text typically consists of an unsegmented stream of words. Parsing and structural analysis are practically indispensable for the retrieval on the basis of higher meaningful linguistic units and for the possibility to present to the user the results in such a format.

The automatically acquired text based index is the link between the disclosure and retrieval modules and supports the retrieval of the stored textual representations and (fragments of) the objects linked to the index terms. The system exploits language as a means to filter and narrow down in several steps the space of potentially relevant target objects. One of the obvious advantages of this stepwise process is that the downloading of condense data objects such as images, video streams or sound tracks can be postponed until there is confirmed evidence that there is a match with the actual information need.

Unlike in most ordinary retrieval systems, the index is also in many other respects not limited to an index based on single words or lemmata. In fact, it is a combination of several indexes, comprising a fuzzy phrase-based index, a lemmatised vector space index and a bibliographic index. Through the phrase based index, users are allowed to query the system by using not only simple keywords, but also complete phrases, such as: "effects of acid rain on forests in the Netherlands". The matching between query text and index can be done via a one-run fuzzy match that ranks documents on the basis of similarity and number of matching phrases. The incorporation of a vector space index allows a user to improve the initial retrieval results by feeding the most relevant pages back into the retrieval system to get similar documents returned. This mixed approach has been proven to yield a considerable improvement in retrieval performance. Recall profits from the morphological analysis (compound splitting) and fuzzy matching, step-wise retrieval with user interaction and relevance feedback improves precision.

On top of monolingual retrieval, OLIVE supports cross-language information retrieval (CLIR), following also the approach developed within TWENTY-ONE. For example, videos with a German soundtrack are made accessible via queries in any of the languages French, English, Dutch and German. For this aspect of the retrieval functionality two options are developed: off-line document translation using commercial MT-software (specifically the LOGOS MT-server), and on-line query translation. Which option is offered, depends mainly on the resources available (e.g. translation dictionaries) for each language pair.

In order to evaluate the viability of information retrieval from automatically generated transcriptions, the retrieval precision from both machine and human created transcripts on a small set of audio and video documents was measured. This data, used in the TREC-7 SDR track, contains approximately 100 hours of radio and television broadcast news. Using the LIMSI speech recogniser and the TNO information retrieval system, the results obtained on this data with the machine transcripts (average precision of 0.495) are pretty comparable to those obtained with the human transcripts (average precision of 0.524).

3.3 Inherent Limitations

Obviously, 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 project 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

OLIVE is funded by the European Commission under the Telematics Application Programme in the sector Language Engineering, which now turned into the Human Language Technology action line. The project

(LE4-8364) started in April 1998 and will last until 2000. The results thus far comprise a detailed overview of user requirements, a detailed functional design for the demonstrator, an update of the data capture tools developed within Pop-Eye and a so-called lab model, which offers the proof of concept for speech-based video retrieval. This lab model contains a limited amount of digitised video material from an American English news show with a variety of speakers (anchor man, studio guests and people calling in from outside the studio). The sound track has been transcribed by the recognition tools for American English from LIMSI developed previously[1, 2]. The resulting transcripts have been indexed by the disclosure modules, and translated with commercial MT-Software (LOGOS). Queries can be submitted in French, German and English, and the system returns the relevant phrases plus the links to the relevant fragments which can be viewed with a Real-Video plug-in.

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 NOB, a large service provider for broadcasting and TV productions (Hilversum, Netherlands). Technology development and system implementation involve: TNO-TPD (Delft), the project co-ordinator supplying 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 natural language technology, LIMSI-CNRS (Orsay, France) and Vecsys SA (Les Ulis, France) developing and integrating the speech recognition modules, respectively.

More information about Olive, the lab model and links to other relevant projects such as Twenty-One and Pop-Eye can be found under http://twentyone.tpd.tno.nl/olive.

References

1. J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, Feb. 1997, pp. 56-63.
2. J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, Sydney, Nov. 1998, pp. 1335-1338.
3. F.M.G. de Jong "Twenty-One: a baseline for multilingual multimedia retrieval", *Proceedings of the 14th Twente Workshop on Language Technology (TWLT-14)*, University of Twente, 1998, pp. 189-194.
4. W.G. ter Stal. J-H Beijert, G. de Bruin, J. van Gent, F.M.G. de Jong, W. Kraaij, K. Netter and G. Smart, "Twenty-One: Cross-language disclosure and retrieval of multimedia documents on sustainable development," *Journal of Computer Networks and ISDN Systems* Vol. 30, Elsevier, 1998, pp. 1237-1248.
5. Hiemstra, D. and W. Kraaij, "Twenty-One at TREC-7: Ad-hoc and Cross-language track," *Proc. of the Seventh Text Retrieval Conference TREC-7*, NIST Special Publications, 1999.
6. G. Jones, J. Foote, K. Sparck Jones and S. Young, "The video mail retrieval project: experiences in retrieving spoken documents," Mark T. Maybury (ed.) *Intelligent Multimedia Information Retrieval*, AAAI Press, 1997.