Language-Based Multimedia Information Retrieval

Franciska de Jong
University of Twente
Dept. of Computer Science/CTIT
P.O.Box 217
7500 AE Enschede, Netherlands
fdejong@cs.utwente.nl

Jean-Luc Gauvain
LIMSI-CNRS
B.P. 133
91403 Orsay Cedex, France
gauvain@limsi.fr

Djoerd Hiemstra
University of Twente
Dept. of Computer Science/CTIT
P.O.Box 217
7500 AE Enschede, Netherlands
hiemstra@cs.utwente.nl

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

Abstract

This paper describes various methods and approaches for language-based multimedia information retrieval, which have been developed in the projects POP-EYE and OLIVE and which will be developed further in the MUMIS project. All of these project aim at supporting automated indexing of video material by use of human language technologies. Thus, in contrast to image or sound-based retrieval methods, where both the query language and the indexing methods build on non-linguistic data, these methods attempt to exploit advanced text retrieval technologies for the retrieval of non-textual material. While POP-EYE was building on subtitles or captions as the prime language key for disclosing video fragments, OLIVE is making use of speech recognition to automatically derive transcriptions of the sound tracks, generating time-coded linguistic elements which then serve as the basis for text-based retrieval functionality.

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 (audio-)visual data.

Without any doubts, image and video processing have made enormous progress over the past years, for example, in the areas of analysis of low level or maybe even higher level image features and of segmentation of continuous video material. Low level analysis of texture and colour histograms can already form the basis for retrieving images of similar kinds. The detection of movements can help to identify and retrieve sequences in which movements occur, which already by itself can be an application, as in the case of automatic surveillance. The recognition of simple shapes can be a first step in the direction of recognising and identifying certain objects. The progress that has been made in the area of segmentation, i.e., the identification of shots or even scenes, by now also allows to identify more complex shot boundaries, as they are found in wipes, fading or other kinds of transitions.

Segmentation is relevant not only for identifying the extension of a coherent sequence, but also and above all for breaking down continuous material into static or spatial dimensions by representing such shots through one single picture. In the simplest case, this is typically achieved by key frame extraction, where one significant frame is taken to represent one shot. However, there are again also much more complex types of representation building on mosaicing, where the camera sweep of an entire shot is put together into one static panoramic picture.

Besides all this progress, there appear to be still two major unsolved problems in the broad scale indexing and retrieval of video material on the basis of the described technologies, viz., (a) image and video-processing is still far away from understanding the content of a picture in the sense of a knowledge-based understanding, and (b) there is no effective query language (in the wider sense) for searching image and video databases. In a certain sense, these two problems are really two sides of the same issue: leaving aside all the philosophical intricacies associated with this question, there is simply no escape from the fact that human language plays a central role in representing, expressing and also processing knowledge. One of the crucial consequences of this state of the art is, of course, that so far for image and video obects, indexing and retrieval is still practically impossible without the intervention of interpretation by a human who understands the audio-visual content and describes it in the format of a human language, which then can serve as the platform for language-based search and retrieval.

To tackle the problem of automatic disclosure and retrieval of audio-visual material, the two EU-funded projects POP-EYE[1] and OLIVE [2]are therefore trying to exploit the linguistic information associated with such data. They are both building on human language as the media interlingua, making the assumption that, as long as there is no possibility to carry out both a broad scale recognition of visual objects and an automatic mapping from such objects to linguistic representations, the detailed content of video material is best disclosed through the linguistic content associated with the images. In the case of POP-EYE, which was launched at a time when practically only written material could be reliably processed, the prime linguistic data were subtitles (close or open captions) associated with videos. OLIVE, more or less a follow-up project, extends the range or variety of linguistic data and focuses on speech technology processing of the sound track, but also takes into account other linguistic material associated with video documents. Both projects make the assumption that while we have no access to the visual content directly, one should at least make use of the linguistic content of video data, which may also provide a direct or indirect reflection of the visual content. Clearly this cannot provide a universal solution for the problem of automatic disclosure and retrieval, but at least it can contribute to the automatic capturing of as much of the information as is possible by the state of the art.

The main objective of these projects is thus to develop a radio and video archiving and retrieval tool that will facilitate efficient access to large libraries of audio-visual material. In order to allow a detailed retrieval, the indices that are built from the associated linguistic material are related to time-codes whenever this is possible, i.e., they point to particular frames or shots in the video rather than to the video as a whole. While subtitles themselves already provide a time-coded textual basis, which only has to be indexed appropriately, such a textual basis has to be created in the case of spoken input material through automatic speech recognition. Thus, in the OLIVE project a prototype is developed and tested 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. This process is complemented by employing various alignment techniques for drawing into account other textual material, which is not time-coded yet, but which can be brought into a relation with the time-coded material. For the retrieval part, tools are developed which support users in searching for material via natural language queries, including cross-lingual access based on offline machine translation of the archived documents or alternatively online query translation.

The consortia are comprised of users and technology providers and integrators. The primary users are broadcast organisations (ARTE, BRTN, SWR, and TROS), a national audio-video archive (INA) and

---

[1]Pop-Eye is a EU-funded project within the Telematics Application Programme, sector Language Engineering (LE-4234). Duration: 1997-1998.

[2] Olive is a EU-funded project within the Telematics Application Programme, sector Language Engineering (LE-8364). Duration: Spring 1998- Summer 2000.

a large service provider for broadcasting and TV productions (NOB). Technology providers include TNO, the University of Twente and DFKI for retrieval technology and natural language processing, LIMSI-CNRS for speech recognition technology, the University of Tübingen for evaluation, and two industrial companies, VECSYS and VDA, for integration and exploitation.

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 projects and a short description of the demonstrators that have been built.

2 User Needs

The prime target users of the projects are professionals with an interest in an efficient, detailed and direct access to their video archives. 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 feauture. 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.[3] 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 vast majority of material cannot be disclosed on this basis at all.method can only be applied to selected productions, and that the vast majority of material cannot be disclosed on this basis at all.

The projects aim 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 systems also provide access to the digitised video material through network technology, specifically web browsing. This answers 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 is 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 behind these search environments for video material is that the user can narrow down his search by first inspecting information in the form of index terms, text passages, transcriptions or subtitles, story-boards or sequences of key-frames, in order to finally focus in on the actual condense data objects such as video sequences.

---

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

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 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 (developed in OLIVE for 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 alignment 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.

The retrieval functionality is building on some of the core functions of a search engine whose very first foundations were developed in the Twenty-One project (http://twentyone.tpd.tno.nl/). This search engine is described in more detail below. To support cross-lingual search and retrieval, different approaches are pursued, such as employing translation technology for offline document translation where the translated documents serve as the basis for indexing, and for online translation of query terms where the translated query is then matched against an index build from the text in the original language.

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, which can be used to guide the search during recognition, what can be qualified as informed speech recognition. The time-codes produced by the speech recognizer can be used to align the hypothesized transcription with the text of the original document.

Fully automatic transcription is provided by the state-of-the-art speech recognizer developed at LIMSI [hub4y97, cacm00]. This recognizer makes use of continuous mixture density HMMs for acoustic modeling, combined with a 65k word four-gram language model. Decoding is carried out in multiple passes, incorporating cluster-based test-set acoustic adaptation. Confidence scores are associated with each hypothesized word to allow further processing steps to take into account the reliability of the candidates.

Prior to word recognition, the acoustic signal is partitioned into homogenous segments, and appropriate labels are associated with the segments [icslp98]. 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 [hub4y97] 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, which aims to reduce the mismatch between the models and the data, a critical step for generating accuracte word graphs.

Taking advantage of the corpora available through the LDC, the speech recognizer [hub4y97,icslp98] was first 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 about 20 hours of broadcast data collected in OLIVE for each language, LIMSI has ported its American English system to French and German.

Experiments with 600 hours of unrestricted broadcast news data indicate that word error rates around 20% are obtained for American English (measured on a representative 10 hour subset). Experiments on about 3 hours of French and German broadcast news data indicate that the word error rates are slightly higher (about 25%), 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 perfect word recognition is not 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 given the current level of performance of information retrieval techniques, recognition errors do not add new problems for the retrieval task [mayb97,cacm00].

3.2 Alignment of non-time-coded text

Besides time-coded texts coming from subtitles or speech recognition, for many programs there is also a rich source of non-time-coded texts available. These texts can be the direct or indirect result of the production itself, such as scripts, autocue files or manual transcriptions produced for translation, but the texts can also come from other sources, as for instance press releases or reviews.

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 automatic speech recognition. The OLIVE alignment module derives a time-coding for these texts by automatically aligning them with the time-coded output of speech recognition. The results of alignment can be used to replace the imperfect output of automatic speech recognition or otherwise to complement the output of speech recognition. Users may want choose the former option if the non-time-coded text is a manual transcription of the program and the latter option for related texts.

For the development of the alignment module, several statistics- and heuristics-based approaches were tested using, for example, character frequencies, word frequencies and stop lists [sluis00]. Preliminary tests with alignment could not be performed on the actual output of speech recognition, because this data would only be available near the end of the OLIVE project. In lack of this data, alignment was tested using time-coded closed caption subtitle files of news broadcasts provided by one of the users.

In a first evaluation, autocue files referring to the same programs as the subtitle files were aligned to the time-coded subtitle files. The autocue files in these tests serve as near perfect manual transcriptions of the program that could be used to replace the results of speech recognition. The evaluation showed that an average performance of 98 % precision and 51 % recall on manually aligned test data. The use of additional heuristics that take into account the successful alignments for surrounding sentences could improve the recall up to 81 % reducing the precision only to 95 %.

In a second test the time-codes were removed from closed caption files belonging to news programs that were broadcast on the same day as the programs of the time-coded subtitle files. The resulting non-time-coded subtitle files serve as related material. They belong to programs that cover a lot of the same news events, but possibly in a different order and possibly with one or two items that should not be aligned at all. On these data the basic alignment algorithm achieved a precision of 75 % and a recall of 70 % on average. The use of additional heuristics that take into account surrounding

alignments improved the performance results considerably to a precision of 76 % and a recall of 92 % on average.

Although the pilot evaluations were not done with the actual output of speech recognition, it is nevertheless quite likely that they are a reliable indication of the alignment performance in the final OLIVE system.

3.3 Indexing and Retrieval

The retrieval functionality employed builds on technology developed within the Twenty-One project which produced 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. [trec99].

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 [twlt98] and [isdn98].

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 retrieval module contains a search kernel supporting several query modes and query 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 weighted lemma-based 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 weighted lemma-based index, that uses a successful new probabilistic term weighting algorithm developed at the University of Twente [trec99], 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 (including 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 Machine Translation (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 Spoken Document Retrieval 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).

3.4 Cross-language Retrieval

On top of monolingual retrieval, the projects described also support cross-language information retrieval (CLIR), where cross-language retrieval means that information originally available in one language is retrieved as a response to a query in another language. The basic options available for CLIR are illustrated in figure 1.

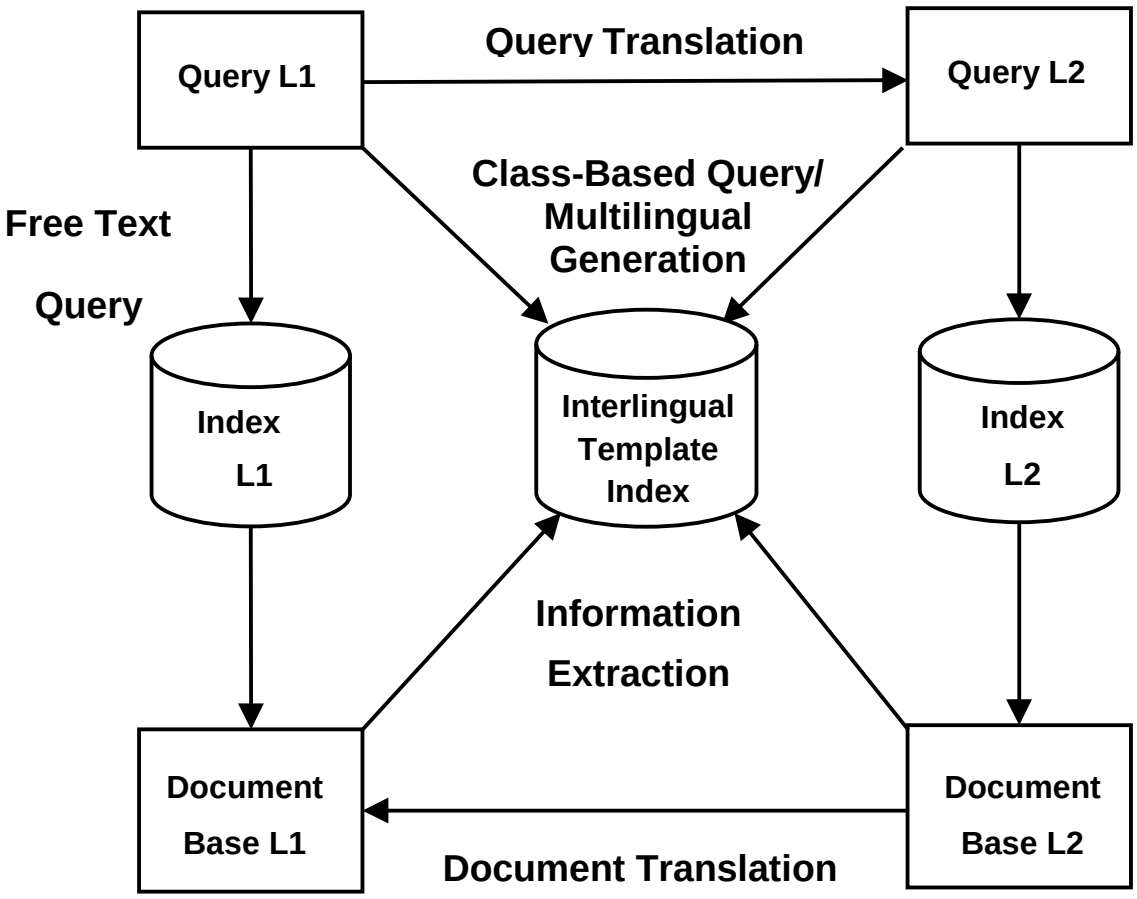


Figure 1 : option for Cross-Language Information Retrieval

The first option we will refer to as offline document translation. To our knowledge this method was first incorporated in a publicly accessible search engine in the Twenty-One project.[4] In this approach, the documents in one language (L2) are automatically translated offline into another language (L1). On this translated document base L1, standard monolingual IR techniques can be applied, i.e., an index for L1 can be created which can then be accessed via a query in the language L1. This document base could of course also be a mixed document base, which contains to the translated documents also original documents of L1.

The second approach is commonly referred to as online query translation. In this case the query in the language L1 of the user is translated into a query in the language L2. This query is then matched against an index created from the original documents in L2, and the original documents are retrieved. One of the first projects to incorporate this approach was the Mulinex project [Mulinex-Reference].

The third option for providing multilingual access builds on information extraction and the construction of language independent representations in the form of interlingual templates or other relational structures. From these language, either textual representations in the different languages are created through automatic multilingual text generation, or the representations have some other direct correspondences in the forms of menu items. Typically, such interlingual representations are queried by means of forms or other structures which can be mapped onto the templates. This approach has

[4] In all of the projects under discussion here, which made use of document translation, i.e., Twenty-One, Mulinex, Pop-Eye, OLIVE, and Mietta, the Translation Server of the LOGOS Corporation has been employed.

been realised for one of the first times in the Mietta project, which provides a combination of class-based querying with free text search (in addition to the crosslingual search facilities based on offline document translation). [cf. Buitelaar/Netter/Xu 1998]

All of these options are realised in one way or the other in the projects under discussion; Pop-Eye was build exclusively on offline document translation, while OLIVE focuses more, but not exclusively on query translation. Another EU funded multimedia retrieval project, Mumis, which is about to be launched, will realise the third approach described. Mumis will realise a detailed disclosure of videos of soccer matches by exploiting again different sources of linguistic information, such as spoken comments transcribed by automatic speech recognition, news paper reports on matches, or other kinds of material. All of this material is submitted to some information extraction process, whose objective is to extract templates or frames from the text which describe certain actions in the game. The extracted information is then stored in a concept-like representation, which can be searched in different languages through direct mappings of concepts onto language specific terms.

Now the question is of course, which of these three options provides the best solution. Unfortunately neither there is neither a clear theoretically nor empirically fully satisfactory answer to this question. In an ideal world, where fully automatic MT works with high precision and for all language pairs, where there are no space and time limits, the document translation approach would most likely be the ideal solution. It requires the least knowledge of the foreign language from the user, as he can both formulate the query in his own language and retrieve the document in the language of his choice independent of the original language. In practice, the solution is not quite as ideal. Not all language pairs that are needed and required are covered by commercial MT systems. Even the high-quality systems still produce only translations which are suitable for understanding the wider content of the original. The approach requires that the documents be translated and the translations stored, if one wants to provide the translations (as an option) to the users. And, which may be most crucial from the retrieval point of view, the MT system determines the retrieval quality. If it mistranslates a term and if this mistranslation is indexed, the user has practically no possibility to get back to the original term and the original meaning.

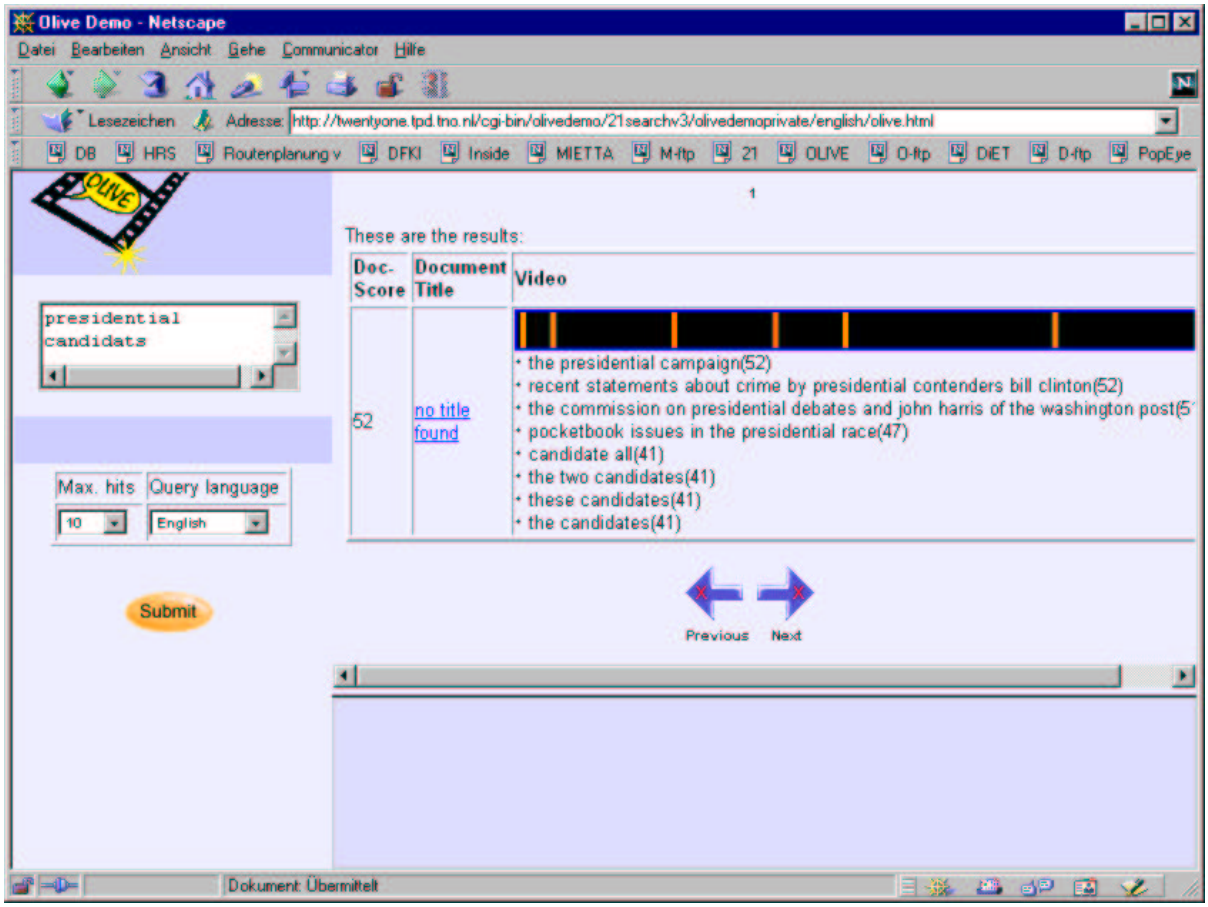
Query translation approaches, which often serve mainly as translation aids to the user, do not have this disadvantage Typically the user is offered a range of possibly translations by the system, from which he can choose the best translation, or even add a translation if he is unsatisfied with all of the options. The biggest disadvantage of the approach is of course, that it requires at least some passive knowledge of the foreign language from the user.

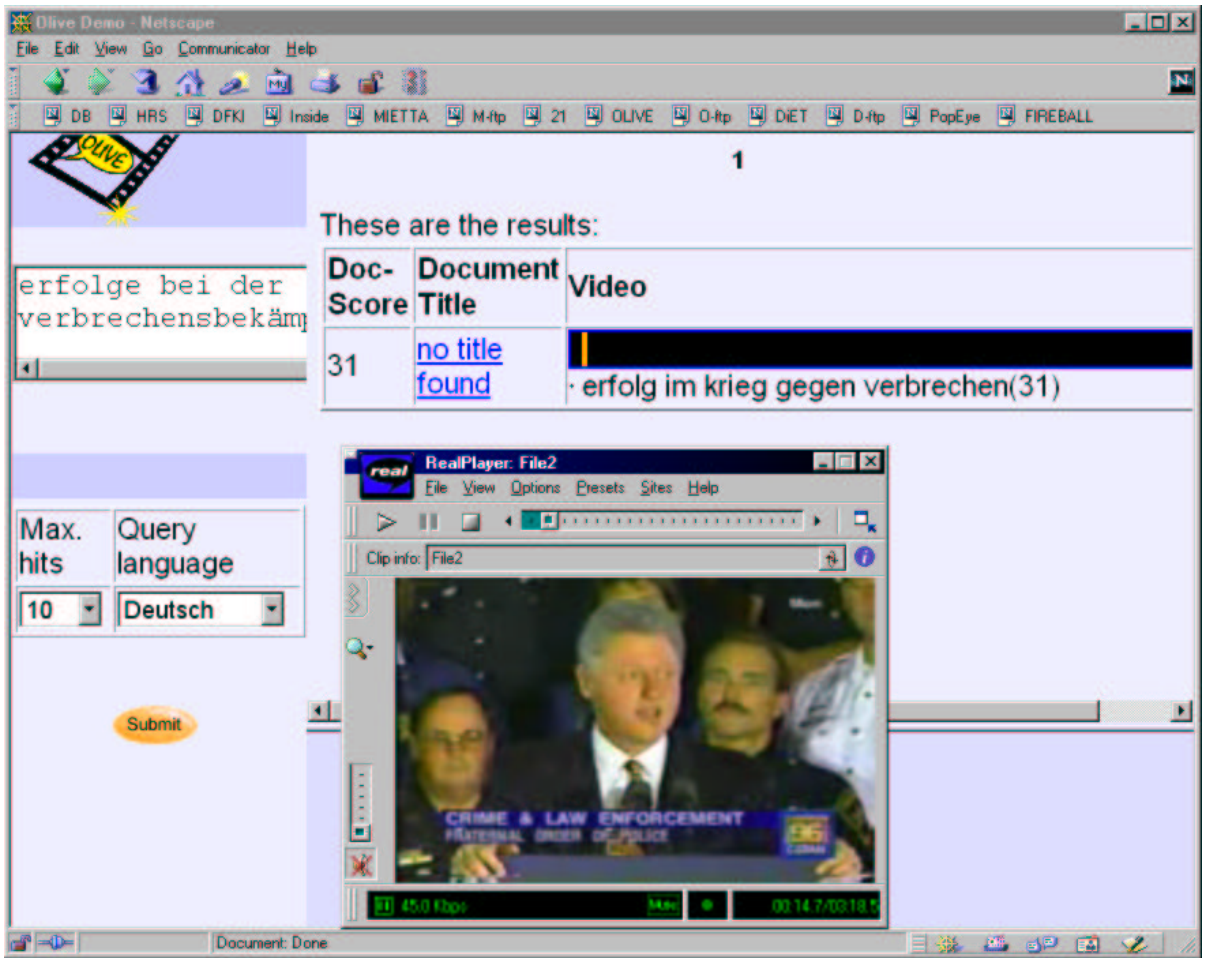
The third approach, using information extraction and multilingual generation, is the best controlled way for providing multilingual information, however it probably also the most restricted one. It is suitable above all for those configurations and contexts, where the information of interest is best represented in a highly structured way, as in a relational database, where the respective structures are highly repetitive, and where not all of the information available is of interest, but only selected parts. Thus video contents, which are very diversified are probably the most unsuitable for this approach, while, for example, soccer games, where the basic types of actions in their essence are fairly limited and also quite repetitive, can be most likely described by fairly standardised representations, which are then also easily mapped onto different languages. Whether this expectation carries out will be one of the interesting outcomes of the Mumis project referred to above.

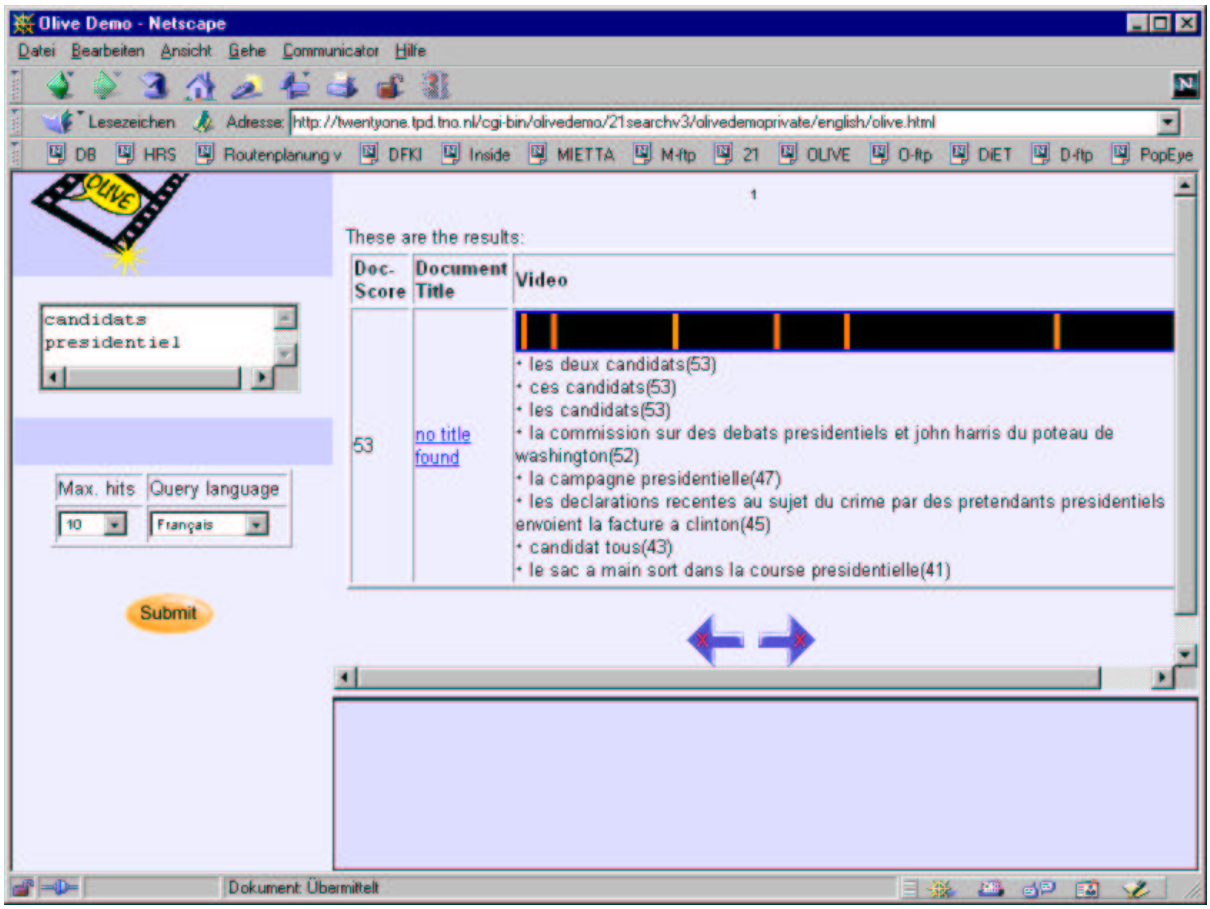
3.5 Inherent Limitations

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 described 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 Screenshots from the OLIVE Lab Model







5 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 [hub4y97,icslp98]. 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

[cacm00] J.L. Gauvain, L. Lamel, & G. Adda (2000), "Transcribing broadcast news for audio and video indexing." In : *Communications of the ACM*, 43(2).

[hub4y97] J.L. Gauvain, G. Adda, L. Lamel & M. Adda-Decker (1997), "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System." In : *Proceedingsof the ARPA Speech Recognition Workshop*, pp. 56-63.

[icslp98] J.L. Gauvain, L. Lamel & G. Adda (1998), "Partitioning and Transcription of Broadcast News Data." In : *Proceedings of ICSLP'98*, Sydney, pp. 1335-1338.

[twlt98] F.M.G. de Jong (1998), "Twenty-One: a baseline for multilingual multimedia retrieval." In ; *Proceedings of the 14th Twente Workshop on Language Technology (TWLT-14)*, University of Twente, pp. 189-194.

[isdn98] W.G. ter Stal. J-H Beijert, G. de Bruin, J. van Gent, F.M.G. de Jong, W. Kraaij, K. Netter & G. Smart (1998), "Twenty-One: Cross-language disclosure and retrieval of multimedia documents on sustainable development." In : *Journal of Computer Networks and ISDN Systems* Vol. 30, Elsevier, pp. 1237-1248.

[trec99] Hiemstra, D. & W. Kraaij (1999), "Twenty-One at TREC-7: Ad-hoc and Cross-language track." In : *Proceedings of the Seventh Text Retrieval Conference TREC-7*, NIST Special Publications.

[mayb97] G. Jones, J. Foote, K. Sparck Jones & S. Young (1997) "The video mail retrieval project: experiences in retrieving spoken documents." In : Mark T. Maybury (ed.) *Intelligent Multimedia Information Retrieval*, AAAI Press.

[sluis00] I.F. van der Sluis & F.M.G. de Jong (2000), "Enriching Textual Documents with Timecodes from Video Fragments." In this volume.