SemanticVox: A multilingual video search engine

Bertrand Delezoide
CEA/LIST, INRIA Futurs
Centre de Fontenay-aux-Roses
BP 6 92265 Fontenay-aux-Roses
33 (0)1 46 54 86 53
bertrand.delezoide@cea.fr

Hervé Le Borgne
CEA/LIST
Centre de Fontenay-aux-Roses
BP 6 92265 Fontenay-aux-Roses
33 (0)1 46 54 85 31
herve.le-borgne@cea.fr

ABSTRACT

In this paper, we describe the SemanticVox project. SemanticVox aims at providing a real link between speech transcription technologies from Vecsys [8] based on LIMSI research [9] and multimedia documents analysis and retrieval technologies from the Multilingual Multimedia Knowledge Engineering Laboratory (LIC2M) of the CEA-LIST [1]. The first application of the project is a cross-lingual automatic video indexing and retrieval system based on speech transcription and video analysis. The two main novelties of the system are: (i) its ability to manage multilingual queries and documents; (ii) its innovative ranking to sort relevant documents based on a deep analysis of the syntax and semantic of the query.

A video of this demonstration is available at http://www.eeng.dcu.ie/~hlborgne/semanticvox.wmv

Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content analysis and indexing – *indexing method, linguistic processing*.

H.3.3 [**Information storage and retrieval**]: Information search and retrieval – *search process, selection process.*

General Terms

Algorithms, Management.

Keywords

Video search engine, automatic indexing, multilingual speech transcription, video segmentation, video classification.

1. INTRODUCTION

With the increasing number of digitized video available on the web, managing and searching in video databases became a very active research area. Two aspects of this research must be considered to understand the whole field of research.

First, a large part of the community agrees on using content-based retrieval from digital video to build systems answering efficiently to a user looking for a video (TREC). However, opinions diverge on the type of content that should be used. The answer usually varies between extracting features from the audio part and the visual one. A textual representation of the spoken content from a video can be obtained through (automatic or manual) speech transcription. Information retrieval from transcripts has received attention because it permits a simplified access to complex semantic content. Image retrieval based on similarity matching or semantic query has also been studied for many years. Most image retrieval systems are based on features such as color, texture and shape that are extracted from the image pixels. However, lots of existing systems aiming at exploring photo databases (e.g photo banks [2]), as well as collaborative photo sharing systems (e.g Flick-R [3]), are based on manual annotation matching.

Second, the need to manage a large amount of video involves an automation of video indexing. Automatic speech recognition can be very useful for video retrieval, even if the transcription is not perfect. It is generally agreed that an error rate better than 35% for speech recognition implies that retrieval performances are only 3 to 10% less than retrieval using perfect transcriptions [4]. As users usually prefer keyword-based searches rather than example-based ones [5], automatic video annotation has also become an active subject of study. Video shot categorization is often approached by computing features (colors, texture, and shape) from the visual signal, which are processed with a classifier engine to infer high-level information about a shot [6].

This paper describes the SemanticVox video search engine system developed by CEA-LIST [1]. There are two main novelties in our system. The first consists of managing multilingual (French, English US, Spanish, German, Arab) queries and documents. The result of the retrieval process is independent of the language used to formulate the query (among the 5 languages) as well as the language of the videos returned (the language of the document can be different from the language of the query). The second novelty consists in the relevance ranking of the documents. It relies not only on the co-occurrence of the terms of the query (this is classic to sort the document) but also on the real syntax and semantic of the query. It includes the management of synonyms as well as named entities recognition (section 3.4).

The remainder of this paper is organised as follows. Section 2 presents a global overview of our system, describing information processing from video indexing to searching. The different modules of the system, namely multilingual speech transcription, automatic shot segmentation/classification and retrieval system are presented in Section 3. And conclusion is given in section 4.

CIVR'07, July 9-11, 2007, Amsterdam, the Netherlands.

2. System overview

The SemanticVox video search engine is a video indexing and retrieving system. It was conceived for managing and searching in large multilingual video databases such as news archives, documentary banks, conferences or (later) internet sharing systems. SemanticVox supports web-based remote access and has an XML-based architecture that uses a simplified video description scheme. A key objective of our system is to be able to manage large data scale, thus it does not require any manual handling and is fully automatic. The figure 1 presents a global view of the system, from video indexing to searching.

The system indexes every video according to two different channels. On the one hand, the audio speaker segmentation and the multilingual speech recognition give a segmented textual representation of the speech content from the videos (Section 3.1). On the other hand, the visual signal is indexed as follows. First, shots are segmented using a visual segmentation algorithm (Section 3.2). Secondly, some relevant semantic concepts are extracted from shots using image processing and classification (Section 3.2). Finally information from auditory and visual signal is merged based on audio segmentation to produce a semantic content XML-based representation. Text from this representation is then indexed using a multilingual semantic text analyser.

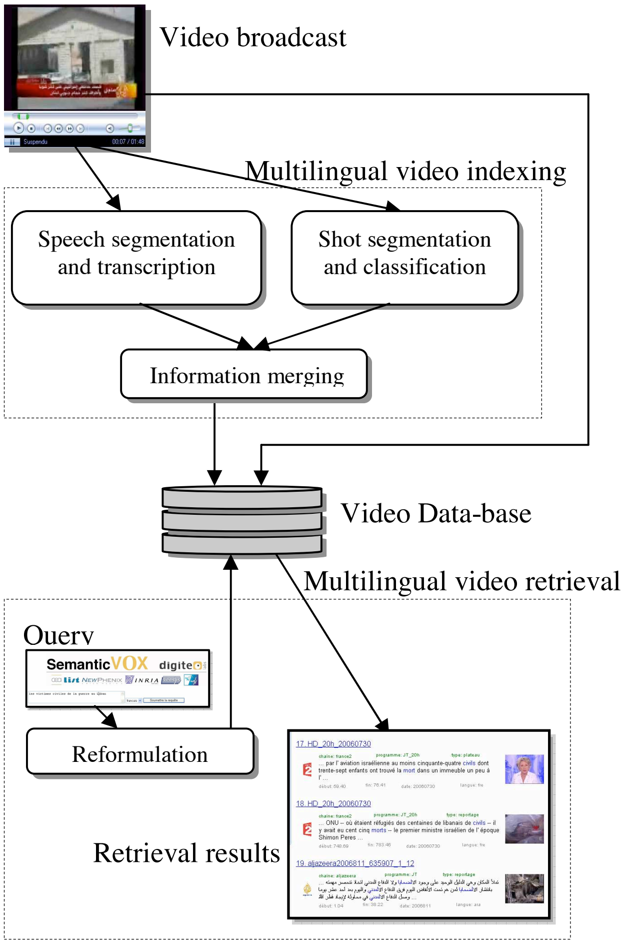
Video broadcast
Multilingual video indexing
Speech segmentation and transcription
Shot segmentation and classification
Information merging
Video Data-base
Multilingual video retrieval
Query
SemanticVOX
Reformulation
Retrieval results

Figure 1. An overview of SemanticVox video search engine.

The retrieving process is based on textual queries entered by users in the web-based interface. First, the query is analyzed with a multilingual text analyser to extract semantic concepts. Secondly, a multilingual reformulation of the concepts is produced using cross-lingual dictionaries. Thirdly, video excerpts containing concepts from the query are retrieved using the textual search engine. Results are shown on a web page in tables presenting the selected documents. Each excerpt is represented by a link to the video, content meta-data, a “representative” image from the video extracted from shot segmentation and the most relevant part of speech extracted from the transcription. When an excerpt is selected by a user, the video page appears. The full transcription from the excerpt is shown and the player starts at the beginning of the excerpt within the video.

3. Modules description

3.1 Automatic speech segmentation and multilingual transcription

The audio processing component of our video retrieval system is composed of two analyzing functions.

First, the audio content is segmented into speech-segments using a GMMs and agglomerative clustering based speech partitioning algorithm, developed by the LIMSI [7]. This algorithm showed significant segmentation results on broadcast news during the NIST evaluation.
Second, speech segments from the multilingual audio signal are transcribed by Vecsys [8] transcription scheme based on LIMSI technologies [9]. This technology, applied to broadcast news in English (US), obtained between 8% and 14% of word error rate at the NIST campaign and supplied transcripts for TRECVid2004 video retrieval evaluation [16].

3.2 Shots segmentation and classification

The standard algorithm we used [10] performs shots segmentation by local maxima detection of an observation function. This function is based on a wavelet transform of color and luminosity from images within the video. The algorithm is optimized to over-segment the signal, nearly every transition is detected as recall is 98 % and some false alarms are spotted as precision is 86 %. We assume that content within a shot is homogeneous. As a consequence, the analysis of a shot can be performed globally rather than for each image within the shot. Thus we chose to extract the representative image from each shot as the one located at the minimum of the observation function within each shot. Semantic concepts are extracted from this image, (we assume that its content contains enough information about the shot to summarize it).

Images are classified using a standard SVM-based classification algorithm [11]. In order for user to access the beginning of an outside broadcast, we chose to first classify shots in stage/outside broadcast. The SVM algorithm shows good performances as 87% of the shots are properly classified (measured on 10 hours of broadcast news). Further treatments such as face recognition or place identification [17] will be considered for the next version of the system.

3.3 Auditory and visual architecture merging

The audio processing component of our video retrieval system produces XML-based representation containing the information related to speech-segments: starting-time, ending-time and transcripts. The visual processing component also produces XML-based representation containing information about shots (start, end, semantic concepts).

Information from audio and visual signals are merged into a single video representation. We define the basic unit of indexing and retrieval as an audio speech-segment. We believe that speech usually conveys most of the information within broadcast news. Then we expect that users retrieving broadcast news are more interested in the speech content of the video than in the visual content. Video retrieval system must then be able to map between visual content and speech-segments that contain the relevant content. A XML file is built on the basis of audio representation from both contents.

3.4 Textual indexing and reformulation

3.4.1 Linguistic treatment

As input, this component receives a document (text or XML format) and identifies the used language. It produces the structure of the document (decomposition in indexing unit) in a specific format containing, for each unit, the standardized content of the transcribed text and the set of properties to be kept in order to index each unit (file names, dates, time code of the speech-segments, image of reference, etc.).

3.4.2 Reformulation

The functional component of *reformulation* makes possible to find a word, or a set of words, independently of languages used. Although this is not an actual translation, each (set of) word(s) is reduced to a language-independent concept. More precisely, it works as follows. Simple words are translated thanks to cross-lingual dictionaries and are thus no problematic. The most interesting part is that complex concepts can be also identified thanks to a two step process. First, each term is translated, leading to a list of potential composed set of words with ambiguities. To solve these ambiguities, the system seeks every set of words in the test corpus and the most likely is retained. In addition to these two steps, a qualification of the words of the query is made. For instance, it determines whether a given term is a derivation, a synonym, a generalisation or a specialisation of a concept.

The entry of this component is a word (the language has been previously identified) and the type of desired reformulation. It returns a structure describing the various forms of reformulation and translation.

Thanks to this *reformulation*, complex semantic content may be retrieved and direct access is granted. The multilingual search engine authorizes querying and retrieving videos in any of the treated languages. For instance, if one formulates the query “The civil victims during the Lebanon war”, semantic reformulation extends user query to known synonyms (victim=>casualty) and other possible semantic associations (Lebanon=>Beirut, as Beirut is the capital of the Lebanese Republic). Whatever could be the language, the system will automatically determine that the user wants some videos dealing with (i) the concept “victim” (that groups the actual words “casualty”, “dead” and so on); (ii) preferably victims that are “civilians” during a “war” (that is also a concept grouping the terms “conflict”, “battle”,…); (iii) that the war the user is interested to is the “Lebanon war” (by default, the most recent one in all the location relating to Lebanon). Finally, the system will then return the best match according to all these criteria, and will decrease the score when less of these are met.

3.4.3 Retrieval

The retrieval component makes possible to retrieve, from a textual query, all the relevant documents and to classify them according to a set of ordered relevance classes. Information about the location of the occurrences of the terms within the documents is furnished as a part of the returned structure. As input, this component receives a specific format of bag of words resulting from the analysis (concept extraction and reformulation) of the textual query. It provides a specific structure identifying the documents organized in relevance classes with information about localization of the occurrences. These “relevance classes” directly correspond to the richness of the matching between the criterions found by the reformulation and the documents returned.

3.5 Access to video

Access to a specific point in the video is granted by a video server called from the web interface. We used Microsoft Windows Media Services 9 Series; it is an industrial-strength platform for streaming live or on-demand audio and video content over the Internet or an intranet [12]. The Windows Media server is designed to handle busy, congested networks and low-bandwidth connections to client computers that are running Video Player.

4. CONCLUSION

In this paper we described the SemanticVox video search engine system. SemanticVox is a multilingual web-based tool for indexing and retrieving videos over large data-bases.

The mutualisation of multilingual automatic speech transcription from Vecsys and multilingual search engine technologies from the CEA-LIST permits to overcome most limitation of standard video search engines such as YouTube [13] or Blinkx [14]. In particular, speech transcription allows users to access to “what is said” in video even if the transcription is not perfect.

There is considerable potential for improving the schemes described for indexing and retrieving video. Future development direction includes: increasing the number of treated languages; improving the navigation in television news by displaying their underlying structure; producing an abstract of every speech-segment; considering scaling-up the indexing modules in order to analyse continuously news channels as CNN or France24 [15].

5. ACKNOWLEDGMENTS

The authors would like to thank to Vecsys [8] for furnishing the audio transcripts. We’d also like to thank the Agence Nationale de la Recherche (ANR) for funding the SemanticVox project within the scope of the science park Digiteo Labs.

6. REFERENCES

[1] CEA-LIST, Laboratoire d'Intégration des Systèmes et des Technologies. http://www-list.cea.fr

[2] Corbis. Royalty free & rights managed search over 500000 quality stock photos. *http://www.corbis.com*

[3] Flick-r. Online photo management and sharing application. *http://www.flickr.com*

[4] Hauptmann, A., Rong Jin, Tobun D. Ng. *Video retrieval using speech and image information*. Proc. SPIE Vol. 5021, p. 148-159, Storage and Retrieval for Media Databases 2003.

[5] Markkula, M. and Sormunen, E. *End-user searching challenges indexing practices in the digital newspaper photo archive*. Information retrieval, 1:259–285, 2000.

[6] Szummer, M. and Picard, R.W. *Indoor-Outdoor Image Classification*, IEEE International Workshop on Content-Based Access of Image and Video Databases, ICCV '98, 1998.

[7] Gauvain, J.L., Lamel, L., Adda, G. *Partitioning and Transcription of Broadcast News Data*, Proc. ICSLP 98, 5, pp. 1335- 1338, Sydney, Dec. 1998.

[8] Vecsys, Speech recognition – Automatic speech treatment. *http://www.vecsys.fr*

[9] Gauvain, J.L., Lamel, L. and Adda, G. *The LIMSI 1999 BN Transcription System*. In Proc. NIST Speech Transcription Workshop, College Park, MD, May 2000.

[10] Josserand, P. *Détection de transitions à l'intérieur d'une séquence vidéo en vue de son indexation*. Master's thesis, Université du Littoral de Calais, 2000.

[11] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, USA, 1995.

[12] *Windows Media Services 9 Series product information*. *http://www.microsoft.com/windows/windowsmedia/forpros/serve/prodinfo.aspx*

[13] YouTube - Broadcast Yourself. Hosts user-generated videos. Includes network and professional content. *http://www.youtube.com*

[14] Blinkx. Search for video and audio clips, using standard keyword and Boolean queries, or conceptual search. *http://www.blinkx.com*

[15] France 24. L'actualité internationale en direct 24h/24, 7jours/7. *http://www.france24.com*

[16] TREC Video Retrieval Evaluation. http://www-nlpir.nist.gov/projects/trecvid/

[17] Delezoide, B. *Modèles d'indexation multimédia pour l'analyse automatique de films de cinéma*. Ph.D. Thesis, Université Pierre et Marie Curie, Paris, France, 2006. http://mediatheque.ircam.fr/articles/textes/Delezoide06c/