# Retrieving Video Segments Based on Combined Text, Speech and Image Processing

**HARRIS PAPAGEORGIOU AND ATHANASSIOS PROTOPAPAS**
Institute for Language & Speech Processing
Maroussi, Greece

**THOMAS NETOUSEK**
Sail Technology AG
Vienna, Austria

## INTRODUCTION

This paper describes a multimedia, multilingual and multimodal research system (CIMWOS) supporting content-based indexing, archiving, retrieval and on-demand delivery of audiovisual content. There are several projects, aiming at developing advanced technologies and systems to tackle the problems encountered in multimedia archiving and indexing [8], [9], [10]. CIMWOS [1] (Combined IMage and WOrd Spotting) incorporates an extensive set of multimedia technologies by seamless integration of three major subsystems – text, speech and image processing – producing a rich collection of XML metadata annotations following the MPEG-7 standard. These XML annotations are further merged and loaded into the CIMWOS Multimedia Database. Additionally, they can be dynamically transformed for interchanging semantic-based information into RDF and Topic Maps documents via XSL stylesheets. The CIMWOS Retrieval Engine is based on a weighted boolean model with intelligent indexing components. An ergonomic and user-friendly web-based interface allows users to efficiently retrieve video segments by a combination of media description, content metadata and natural language text. The database is a large collection of broadcast news and documentaries in three languages (English, Greek, and French), while the open architecture allows for more languages to be incorporated in the future.

### Speech Processing Subsystem (SPS)

When transcribing broadcast news we are facing clean speech, telephone speech, conference speech, music, and speech corrupted by music or noise. Transcribing the audio, i.e producing a (raw) transcript of what is being said, determining who is speaking, what topic a segment is about or which organizations are mentioned are all challenging problems due to the continuous nature of the data stream. One speaker may also appear many times in the data. We would also like to determine likely boundaries of speaker turns based on non-speech portions (silence, music, background-noise) in the signal, so that regions of different nature can be handled appropriately. The audio stream usually contains segments of different acoustic and linguistic nature exhibiting a variety of difficult acoustic conditions, such as spontaneous speech (as opposed to read or planned speech), limited bandwidth (e.g telephone interviews), speech in presence of noise, music or background speakers. Such adverse background conditions lead to significant degradation in performance of the speech recognition systems if appropriate countermeasures are not taken. Likewise, the segmentation of the continuous stream into homogeneous sections (homogenous in speaker and/or acoustic/background conditions) poses serious problems. Successful segmentation however, forms the basis for further adaptation and processing steps. Consequently the above mentioned areas, adaptation to the varied acoustic properties of the signal or to a particular speaker and enhancements to the segmentation process, are generally acknowledged to be key areas for research and improvement to render indexing systems usable for actual deployment. This is reflected by the amount of effort and the number projects dedicated to advance the current state-of-the-art in these areas.

All these tasks are being taken care by the Speech processing subsystem (SPS). The SPS comprises of the Speaker Change Detection module (SCD), Automatic Speech Recognition engine (ASR), Speaker Identification (SID) and Speaker Clustering (SC). The ASR engine is a real-time, large vocabulary, speaker-independent, gender-independent, continuous speech recognizer [2], trained in a wide variety of noise conditions encountered in the broadcast news domain.

### Text Processing Subsystem (TPS)

After processing the audio input, text-processing tools operate on the text stream produced by the Speech processing subsystem and perform the following tasks: Named Entity Detection (NED),

Term Recognition (TR), Story Segmentation (SD) and Topic Classification (TC).

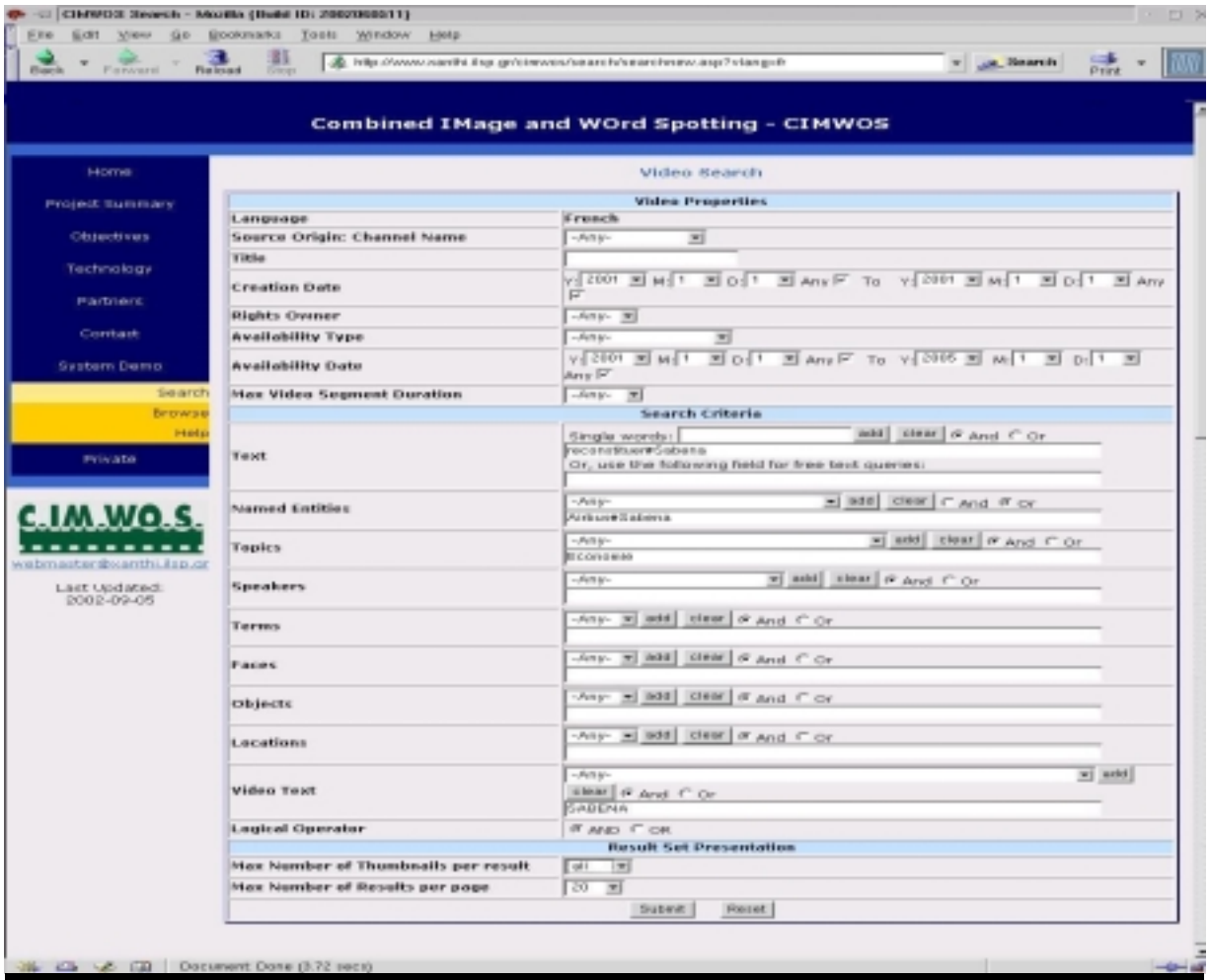



**Figure 1. The CIMWOS query interface**

lemmatization. A lookup module matches name lists and trigger-words against the text, and, eventually, a finite state parser recognizes NEs on the basis of a pattern grammar. Training the NE module consists in populating the gazetteer lists and semi-automatically extracting the pattern rules. For the Greek language, a corpus of 100.000 words per language already tagged with the NE classes was used to guide system training and development.

Another method that has been explored in the project is the construction of a Hidden Markov model that learns to assign a label to every word in the produced text (a label of one of a set of target classes including the class *not-a-class* for any words not pertaining to any specific class of named entity). At decoding time, the most likely sequence of classes given the input text is found and produced as the result of the recognizer.

### Named Entity Detection (NED)

The task of the Named Entity Detection (NED) module is to identify all named locations, persons and organizations as well as dates, percentage amounts and monetary amounts in the text produced by the ASR component. CIMWOS uses a series of basic language technology building blocks, which are modular and combined in a pipeline[3]. An initial finite state preprocessor performs tokenization and sentence boundary identification on the output of the speech recognizer. A part-of-speech tagger trained on a manually annotated corpus and a lexicon-based lemmatizer carry out morphological analysis and

### Term Recognition (TR)

The term recognizer identifies possible single or multi-word terms in the output of the SPS. In the CIMWOS project, a system for automatic term extraction using both linguistic and statistical modelling has been used. Linguistic processing is performed through an augmented term grammar, the results of which are statistically filtered using frequency-based scores. Preliminary testing results have shown that the method was able to locate 62% of technical terminology in a software manual text, compared against a hand-crafted terminology index of it.

### Story Detection and Topic Classification (SD/TC)

Story Detection (SD) and Topic Classification (TC) are both performed using the same set of models. These models are trained on an annotated corpus of stories and their associated topics. The basis of SD and TC is a generative, *mixture-based HMM*. The HMM includes one state per topic and one state modeling general language, that is, words which are not specific to any of the topics. Each state models a distribution of words given the particular topic. After emitting a single word, the model re-enters the beginning state and the next word is generated. At the end of the story, a final state is reached. SD is performed running the resulting models on a sliding window of fixed size, thereby noting the change in topic specific words as the window moves on. The result of this phase is a set of 'stable regions' in which topics change only slightly or not at all. Building on the above located story-boundaries, TC classifies the sections of text according to the set of topic models (including again a model for general language). All technologies used are inherently language independent and of a statistical nature. The models used were trained on several corpora collected and created within the CIMWOS project comprising Greek and French broadcast news (audio and associated transcriptions) and a corpus of French named entities. The modelled inventory of topics is a flat, Reuters-derived structure containing about a dozen of main categories as well as several sub-categories. The annotators had the freedom to add one level of detail to each topic during transcription. Several iterations were needed to arrive at a level of topics which is shallow enough to provide reasonable amounts of training data per topic but still fine-grained enough to allow for flexibility and detail in queries.

### Image Processing Subsystem (IPS)

The image processing subsystem consists of Video segmentation and Key frame extraction (AVS), Face Detection (FD) and Face Identification (FI) spotting persons at any location, scale and orientation, Object Recognition (OR) marking an object in a 'recognition view', given knowledge cumulated from a set of previously seen 'learning views' and Video Text recognition (TDR) detecting and recognizing close captions in images and video.

### Video Segmentation (AVS)

A video sequence consists of many individual images which are called frames and are generally considered as the smallest unit we are concerned with, when segmenting a video. A uninterrupted video stream generated by one camera is called a *shot* (for example, a camera following an airplane, or a fixed camera viewing the 8 pm news presenter). A *shot cut* (or transition) is the point at which shots change within a video sequence.

The goal of video segmentation is to partition the raw material into shots, by detecting shot cuts. For each shot a few representative frames are selected, referred to as *keyframes*, each representing a part of the shot called *subshot*. The subdivision of a shot into subshots occurs when, for example, there is an abrupt camera movement or zoom operation, or when the content of the scene is highly dynamic such that a single keyframe no longer suffices to describe the whole content. Keyframes contain most of the static information present in a shot, so face recognition and object identification can focus on keyframes only. Frames within a shot have a high degree of similarity. In order to detect shot cuts and select keyframes, we have developed methods for measuring the differences between consecutive frames and applying adaptive thresholding on motion and texture cues.

### Face Detection (FD) and Identification (FI)

The FD and FI modules associate faces occurring in video recordings with names. In spite of expending research in face recognition, a lot of problems are still open, particularly in uncontrolled environments, because of lighting, facial expressions, background changes and occlusion problems (glasses or hair for example). One of the most important challenges in face recognition is to distinguish between intra-personal variations (variations in appearance of the same person due to different expressions, lighting, etc.) and extra-personal variations (variations in appearance between persons).

Both modules, FD and FI are based on SVM models trained on a set of an extensive database of facial images with a large variation in pose and lighting conditions [4]. Additionally, a semantic base has been constructed consisting of important persons that

the FI module should identify. During identification, images extracted from keyframes are compared to each model. The decision stage uses the resulting scores from the comparison to either identify the face or to reject it.

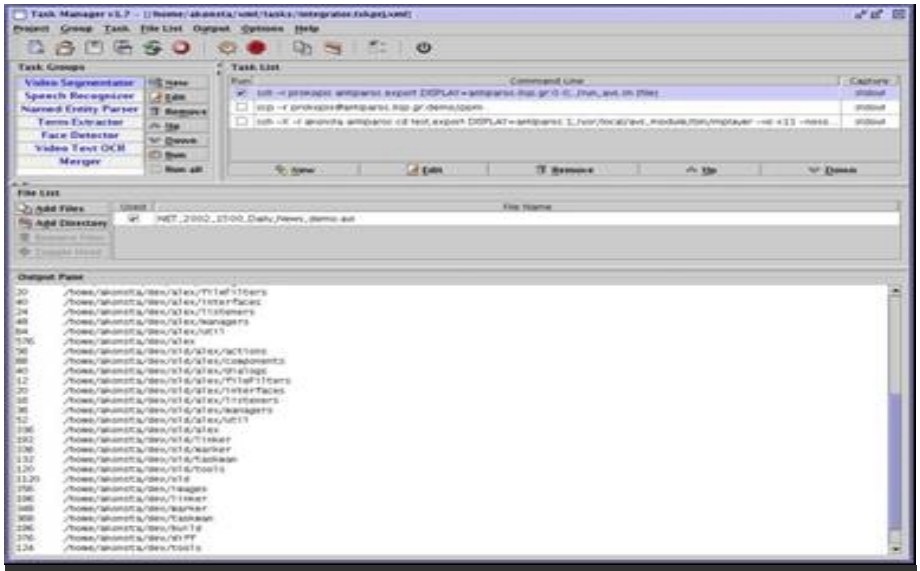


Figure 2. CIIMWOS integration console

## Object Recognition (OR)

Object Recognition (OR) is used to spot and track pre-defined objects of interest. In CIMWOS, the object's surface is decomposed in a large number of regions automatically extracted from the images (small, closed area on the object's surface) [5]. These regions are extracted from several example views (or frames of a movie) and both their spatial and temporal relationships are observed and incorporated in a model. This model can thus be gradually learned as new example views (or video streams) of the object are shown to the system. The strength of this methodology consists fundamentally in two points: First, the regions themselves embed many small, local, pieces of the object appearance at the pixel level, and can reliably be put in correspondence along the example views (or frames). Even in case of occlusion or clutter (eg: background change) in the recognition view, a subset of the object's regions will still be present. The second strong point is that the model, which can automatically be learned from examples, captures the spatio-temporal order inherent in the set of individual regions and requires it to be present in the recognition view. This way the model can reliably and quickly cumulate evidence about the identity of the object in the recognition view, even in cases where only a small amount of recognized regions is found (eg: strong occlusion, difficult illumination conditions, which might make many individual regions hard to spot).

## Video Text Detection and Recognition (TDR)

Text recognition in images and video aims at integrating advanced Optical Character Recognition (OCR) technologies and text-based search, and is now recognized as a key component in the development of advanced video and image annotation and retrieval systems. Unlike low level image features (such as color, texture or shape), text usually conveys a direct semantic information on the content of the video, like a player's or speaker's name, location and date of an event, etc.. However, text characters contained in video are of low resolution, of any color or greyscale value (not always white), embedded in complex background. Experiments show that applying conventional OCR technology directly leads to poor recognition rate. Therefore, efficient location and segmentation of text characters from background is necessary to fill the gap between images or video documents and the input of a standard OCR system.

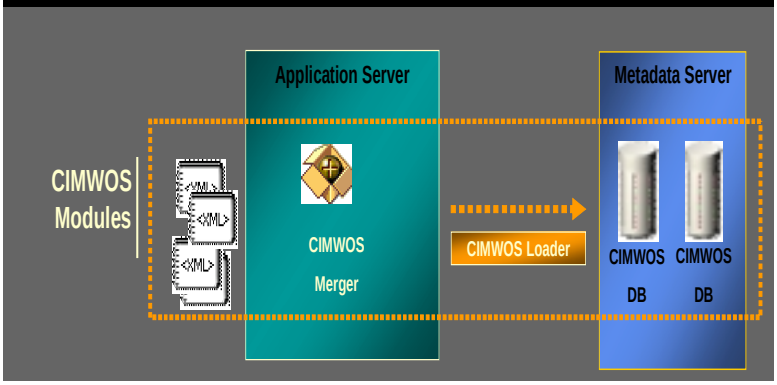
In CIMWOS, the TDR module is based on a statistical framework using state-of-the-art machine learning tools and image processing methods [6]. It consists of four modules: Text detection aims at roughly and quickly finding block of image which may contain a single line of text characters. Text verification based on a SVM model focuses on removing false alarms. Text Segmentation extracts pixels from text images belonging to characters with the assumption that they have the same color/gray scale value. The method exploits a Markov Random Field model and an EM algorithm for optimization. Finally, all hypotheses produced by the segmentation algorithm, are processed by the OCR engine and a string selection is applied, based on a confidence value computation that uses character recognition reliability and a simple bigram language model.

## INTEGRATION ARCHITECTURE

All processing modules in the corresponding three modalities (Audio, Image and Text) converge to a textual XML metadata annotation scheme following the MPEG-7 descriptors. These XML metadata annotations are further processed, merged and loaded into the CIMWOS Multimedia DataBase. The whole architecture is depicted in Figures 2 & 3.

The merging Component of CIMWOS is responsible for the amalgamation of the various XML annotations and the creation of one self-contained object that is compliant with the CIMWOS Database. Additionally, the resulting object can be dynamically transformed for interchanging semantic-based information into RDF (http://www.w3.org/RDF/ ) and Topic Maps (http://www.topicmaps.org/xtm/1.0/) documents via XSL stylesheets (http://www.w3.org/Style/XSL ).

Figure3. CIMWOS integration architecture




## ARCHIVING AND RETRIEVAL

Because of its size, a video clip can take a long time to move from one location to another, such as from the digital video library to the user. Likewise, if a library consists of only 30-minute clips, when users check one out, it may take 30 minutes to determine whether the clip meets users' needs. Returning a full half-hour video when only one minute is relevant is much worse than returning a complete book when only one chapter is needed. With a book, electronic or paper, tables of contents, indices, skimming, and reading rates permit users to quickly find the chunks they need. Since the time to scan a video cannot be dramatically shorter than the real time of the video, a digital video library must be efficient at giving users the material they need. To make the retrieval and viewing of information faster, the digital video library will need to support partitioning video into small-sized clips.

In CIMWOS, the retrieval engine is based on a *weighted boolean* model equipped with intelligent indexing components. The basic retrieval unit is the passage. The passage has the role of a document of a standard retrieval system. The passage is indexed on a set of textual features: words, terms, named entities, speakers and topics. Each passage is linked to one or multiple shots, and each shot is indexed on another set of textual features: faces, objects and video text (caption or scene). Via the linking of passages and shots, each passage is finally assigned a broader set (union) of features that will be used for retrieval. All the features are textual, no image or audio conceptual feature is supported (i.e. motion parameters or similarity between histograms). Each passage is represented as a set of features and retrieval of passages is performed based on computing similarity in the feature space. The results are ranked based on he computed similarity values (Fig. 4).

The retrieval procedure is the following: first, passages are filtered on the basis of the advanced features in order to reduce the search space of the free query. If no such features have been selected, the search space remains unchanged. Next, a boolean-based matching operation takes place in order to compute the similarity between the query and each passage of the possibly reduced space. Finally, passages are ranked based on this similarity.

## CONCLUSION

CIMWOS has developed an integrated environment for accessing multimedia digital content providing advanced searching and browsing capabilities. We have conducted research on emerging technologies for multimedia data processing, indexing and retrieval, and we have embedded them in a video library system that covers broadcast news and documentaries in English, French and Greek. The CIMWOS Digital Video Library uses intelligent, automatic mechanisms that provide full-content search and retrieval from a large (scaling to several hours) online digital video library. The tools that have been exploited during the project can automatically populate the library and support access to it. The adopted approach uses combined speech, language and image understanding technology to automatically transcribe, segment and index the video. Retrieval is text-based, performed on the

material that results from image analysis, speech recognition and natural language processing of the transcribed text. The information that is produced by the speech, image, and text components is suitable for intelligent indexing and retrieval so that a user can find a video segment of interest within the digital library.

## Acknowledgments

This material is based on work supported by the EU Information Society Technologies (IST) program under contract IST-1999-12203.

## References

1. EU IST project CIMWOS (http://www.xanthi.ilsp.gr/cimwos/default.html)
2. F. Kubala, J. Davenport, et al. (1998), "The 1997 BBN BYBLOS System applied to Broadcast News Transcription", Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne VA, Feb. 1998.
3. Demiros, I., Boutsis, S., Giouli, V., Liakata, M., Papageorgiou, H., Piperidis, S., (2000) "Named Entity Recognition in Greek Texts", Proceedings of Second International Conference on Language Resources and Evaluation-LREC2000, 31 May- 2 June 2000, Athens, Greece, 1223-1228.
4. Fabien Cardinaux and Sebastien Marcel (2002), "Face Verification using MLP and SVM", in "XI Journees NeuroSciences et sciences pour l'Ingenieur (NSI 2002)", 2002.
5. T. Tuytelaars, A. Zaatri, L. Van Gool and H. Van Brussel (2000), "Automatic Object Recognition as part of an Integrated Supervisory Control System", in IEEE Conference on Robotics and Automation, ICRA00, pp.3707-3712, 2000.
6. D. Chen and H. Bourlard and J-Ph. Thiran (2001), "Text identification in complex background using SVM", Proc. of the Int. Conf. on computer vision and pattern recognition, pp. 621-626, Dec. 2001.
7. J.M. Odobez and D. Chen (2002), "Robust Video Text Segmentation and Recognition with Multiple Hypotheses", Proc. of the ICIP, Sep. 2002.
8. Wactlar, H., Olligschlaeger, A., Hauptmann, A., Christel, M. "Complementary Video and Audio Analysis for Broadcast News Archives", Communications of the ACM, 43(2), pp. 42-47, February, 2000
9. Michael R. Lyu , Edward Yau , Sam Sze. "Video and multimedia digital libraries: A multilingual, multimodal digital video library system", In Proc. Of the 2nd ACM/IEEE-CS joint conf. On Digital Libraries, July 2002, pp.145-153.
10. Sankar A., Gadde R.R. and Weng F. "SRI's broadcast news system – Toward faster, smaller and better speech recognition", In Proc. Of the DARPA Broadcast News Workshop, 1999, pp.281-286

**Figure 4. Details of a retrieved video segment before streaming**