CIMWOS

www.xanthi.ilsp.gr/cimwos

Combined Image and Word Spotting
IST-1999-12203

Public final project report

September 2003

C.IM.WO.S.
Combined Image
& Word Spotting

Introduction – overview

Nowadays a vast amount of information is stored in the form of video, pictures, and audio, which does not lend itself to automated searching. To improve the usability of these invaluable resources, indexing techniques are required, which are currently very expensive and time-consuming, mainly carried out manually by experts. In view of the expansion of the digital television and of video-based communications and related applications, the need for an editor-like tool that allows the user to see/hear, select/modify and search over audio-visual databases becomes indispensable.

CIMWOS provides content-based indexing, archiving, and retrieval of audiovisual content, aiming to promote reuse of multimedia resources. The system uses a multifaceted approach to locate important segments within multimedia material, employing state-of-the-art algorithms for text, speech and image processing.

The objective of the project was to develop and integrate algorithms specially designed for digital audio-visual content, leading to the implementation and demonstration of a practical system for efficient retrieval in multimedia databases. Project work has led to the design and implementation of a system incorporating multimedia technologies in three major subsystems (text, speech, image) producing metadata annotations. These annotations are loaded into a multimedia database, currently a large collection of broadcast news and documentaries in three European languages. Users can search and retrieve video segments by a combination of criteria entered on a web browser.

Facilitating media indexing using automated procedures and employing powerful search algorithms over multimedia content, the CIMWOS system aims to become a valuable assistant in promoting the re-use of existing resources and thus cutting down the budgets of new productions.

Original project objectives

As defined in the technical annex to the project contract, the main objective of CIMWOS was "to develop a system to automatically locate and retrieve text, images, video, and audio from a multilingual audio-visual database performing content-based searches." It was further indicated that up to four European languages would be supported, with special care taken to ensure that system architecture remain open for new languages.

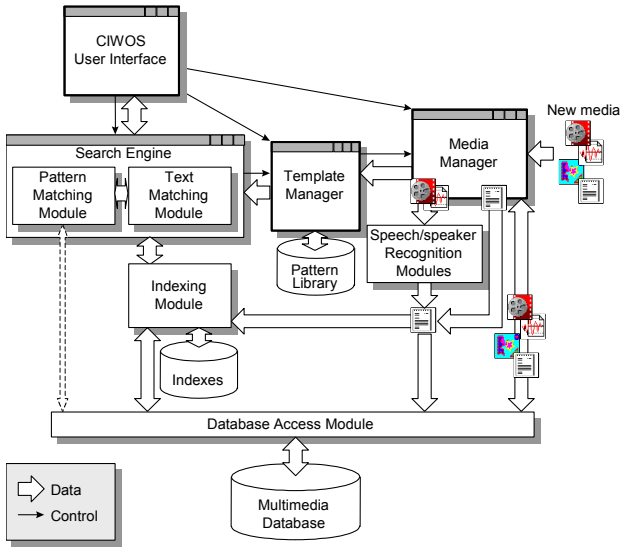
CIMWOS was to use a dual audio and visual approach to locate important clips within multimedia material employing state-of-the-art algorithms for both image and speech recognition. Image processing algorithms would extract features to be employed for pattern matching and speech recognition algorithms would locate keywords in sound-clips and in the soundtracks of the video-clips. This was meant to provide an efficient mechanism for the users to perform their searching tasks, thus enabling more focused and precise searches.

It was specified that CIMWOS would create and maintain a set of indexes to the multimedia contents. During data retrieval, input from both image and speech would be exploited and the principle of combined word and image spotting would be used.

With respect to the speech recognition operation, use of state-of-the-art robust continuous speech recognition algorithms was prescribed to locate words in the videos' soundtracks. An indexing mechanism would store information to associate text fragments (the transcribed speech) with audio fragments, keeping them aligned. As to the image-based retrieval, it was stated that CIMWOS would improve the efficiency of annotation by looking for patterns similar to one specified in a particular image and transfer automatically the annotations to other images for patterns that match.

In addition, CIMWOS was to look for specific classes of objects that can be annotated fully automatically, such as human faces, because of their importance. For faces, automatic transfer of annotations was to be attempted, so that a known face identity would be recognised in other images. Finally, automatic annotation of other object classes was also desirable.

With respect to searching, it was stated that multiple searching criteria could be defined simultaneously, and partial results of each separate search would be combined using predefined logical operators. To preserve bandwidth and increase responsiveness, results of searches would first be transmitted in a compacted "preview" format.



Project achievements

Technically, the project has achieved, in some cases even surpassed its original objectives, and has followed general technological advances while remaining at the forefront of state of the art. Specifically, a system has been developed incorporating an extensive set of multimedia technologies by seamless integration of three major processing subsystems (text, speech, image) producing a rich collection of XML metadata annotations following the MPEG-7 standard. These annotations are merged and loaded into the CIMWOS Multimedia Database. A user-friendly web-based interface allows users to efficiently search and 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 European languages (English, Greek, and French), while the open architecture allows for more languages to be added in the future.

CIMWOS uses a multifaceted approach to locate important segments within multimedia material, employing state-of-the-art algorithms for text, speech and image processing. The audio processing operations employ robust continuous speech recognition, speech/non-speech classification, speaker clustering and speaker identification. Continuous speech recognition technologies produce transcriptions of what is said, determining who is speaking, or what topic a story is about. Text processing tools operate on the transcription produced by the speech recogniser and perform named entity detection, term recognition, topic detection, and story segmentation. Image processing algorithms extract features to be used for pattern matching and recognition of object classes, including scenes, logos, and certain objects, particularly emphasizing detection and recognition of faces. Text displayed in the images (captions, subtitles etc.) is also detected and recognized.

The system has been tested by two user groups, one in Greece and one in Belgium, with respect to both the performance of the processing components and the usability and efficiency of the search and retrieval system. While weaknesses and points needing further development have been identified, it has been concluded that the system is near a practically usable state and that it offers features not available in commercial applications. Indeed, CIMWOS has met with enthusiasm when presented at conferences oriented for the target market of media and broadcasting, and realistic prospects for further development and commercial exploitation have emerged.

CIMWOS potential

CIMWOS addresses real needs of multimedia producers, archivers, and others who need to monitor or index audiovisual content. Use intentions follow a two-pronged strategy:

- Marketing of individual processing components is foreseen according to individual model maturity, which can be integrated into existing products serving the multimedia sector.
- Pilot installation and use of the full system, including processing, archiving, and search/ retrieval, aims to attract investor interest for full development to commercial grade software.

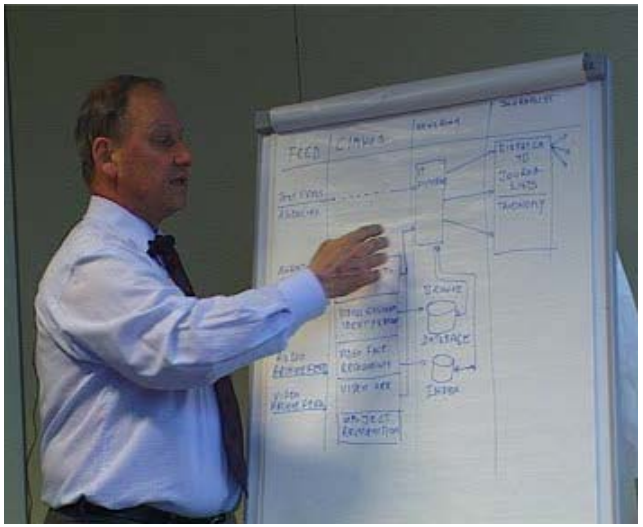
Utilising the vast amounts of information accumulated in audio and video, the CIMWOS system can become an invaluable assistant in promoting the re-use of existing resources, cutting down the budgets for new productions. CIMWOS opens new possibilities and provides enabling technology for novel application areas and services.

With Internet continuing its rapid development and personal computers becoming more and more capable of handling and manipulating multimedia resources on-line content providing services are expected to grow rapidly. However, quality video requires large amounts of storage space and transmission bandwidth, inhibiting on-line circulation via the Internet. CIMWOS offers a viable solution by providing technologies to perform searches locally at the provider site and transmit the results in a compacted "preview" format. Thus, users can download solely the material that suits their specific needs. CIMWOS provides a fast way to assemble information on special areas of interest by exploiting existing archives and thus reducing the time-to-market for new multimedia titles.

CIMWOS can be employed to annotate and give access to large archives and to promote standardisation and unification of multiple depositories thus allowing the rise of multimedia digital libraries at the European level.

CIMWOS can improve the working conditions and the productivity of people involved in media and television, video, news broadcasting, show business, advertisement, and any organisation that produces, markets or broadcasts video and audio programmes, by facilitating common procedures of retrieving audiovisual material. Moreover, it can provide distinct advantages in terms of time-to-market, material reusability, and resource reallocation.

Facilitating access to relevant high-quality information, the CIMWOS system may also become an invaluable assistant for educational purposes delivering the right information at the right time. Being aligned with the ongoing large-scale investment in developing multimedia information systems funded within the European or national framework or by the private sector, the value of CIMWOS may increase further in the future.



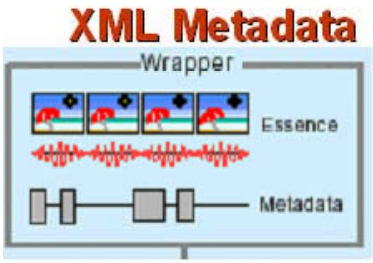
CIMWOS processing structure

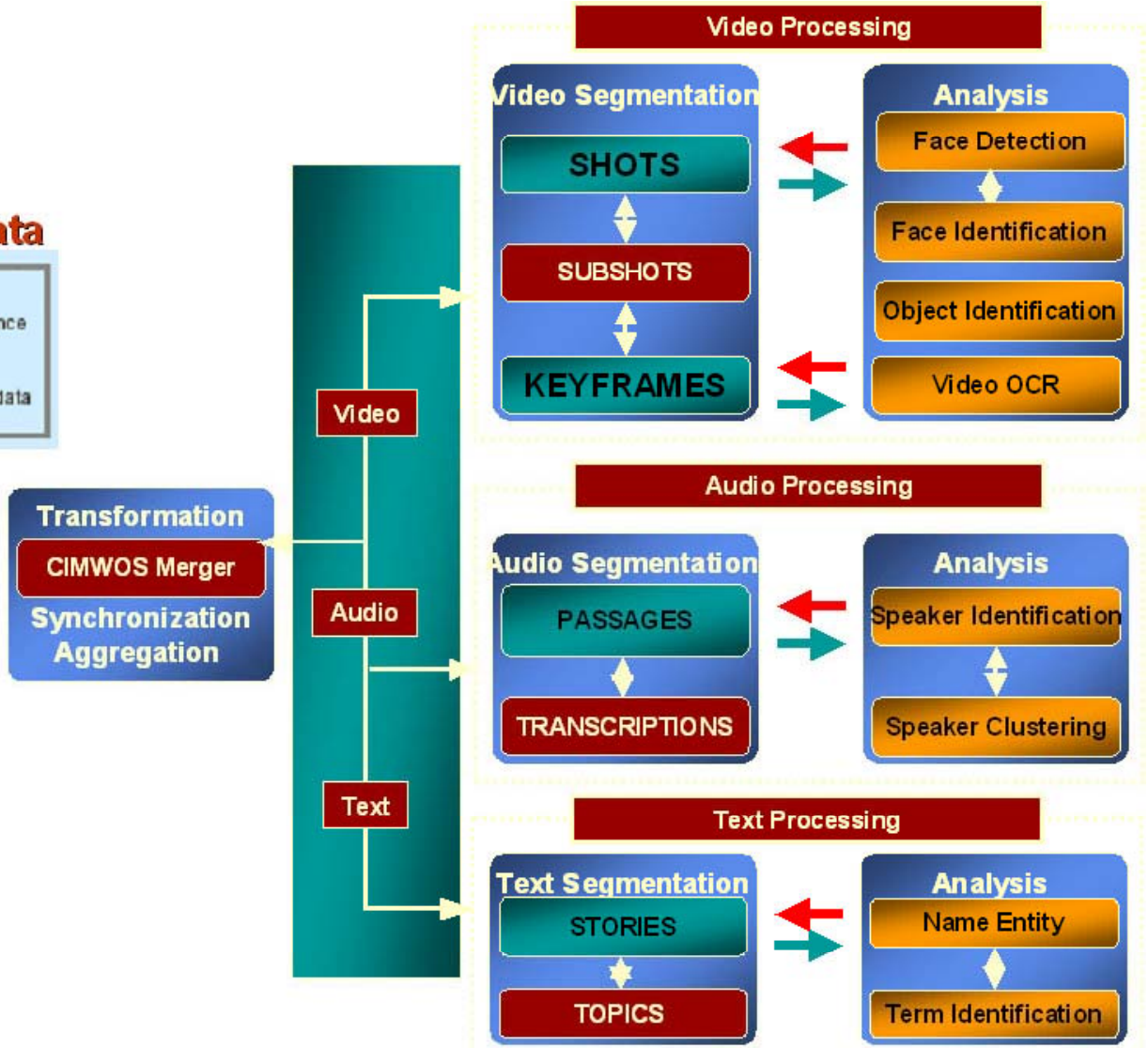
CIMWOS uses a multifaceted approach to locate important segments within multimedia material employing state-of-the-art algorithms for text, speech and image processing.

Audio processing employs robust continuous speech recognition, speech/non-speech classification, speaker clustering and speaker identification.

Text processing tools operate on the text stream produced by the speech recogniser and perform named entity detection, term recognition, topic detection, and story segmentation. Image processing includes video segmentation and key frame extraction, face detection and face identification, object and scene recognition, video text detection and character recognition.

All outputs converge to a textual XML metadata annotation scheme 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.





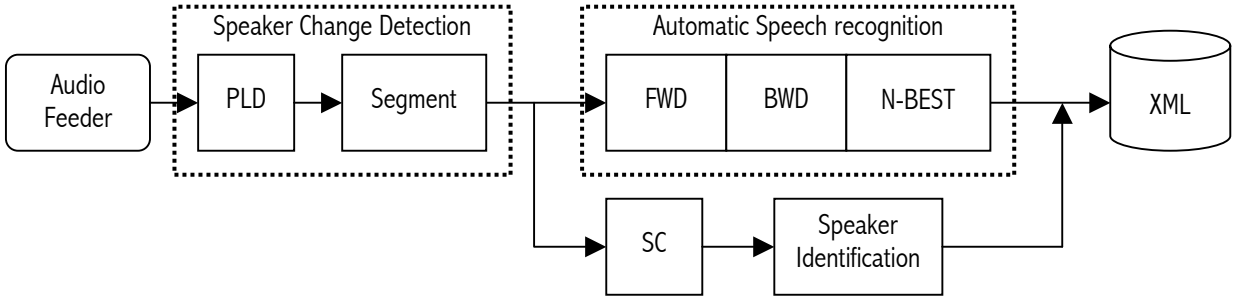
CIMWOS speech technologies

To benefit from the huge potential of information available in broadcast news, efficient methods to process and extract the relevant portions from various digital media must be employed.

Producing a transcript of what is 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. 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 unless appropriate countermeasures are taken. Likewise, the segmentation of the continuous speech stream into homogeneous sections (with respect to speaker, topic, or acoustic/background conditions) poses serious problems. Successful segmentation however, forms the basis for further adaptation and processing steps. 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 of projects dedicated to advance the current state-of-the-art in these areas.

The speech and text processing components in CIMWOS comprise a set of technology components, each one consisting of a set of modules arranged in a pipeline. Within the speech processing component, speaker clustering and speaker identification can be run in parallel to the speech recognition. The input to the overall system is audio, and the final output is a set of XML files containing speech transcriptions, identified speakers, and detected named-entities, terminological units, and categorised stories.

The modular approach allows for maximum flexibility with regard to exchanging technologies on the component as well as the module level. This allows us to benefit from advances made in individual technologies with only minor impact on the architecture of the overall system. Within the CIMWOS project, we will capitalize on this flexibility by using different technologies within each of the components.

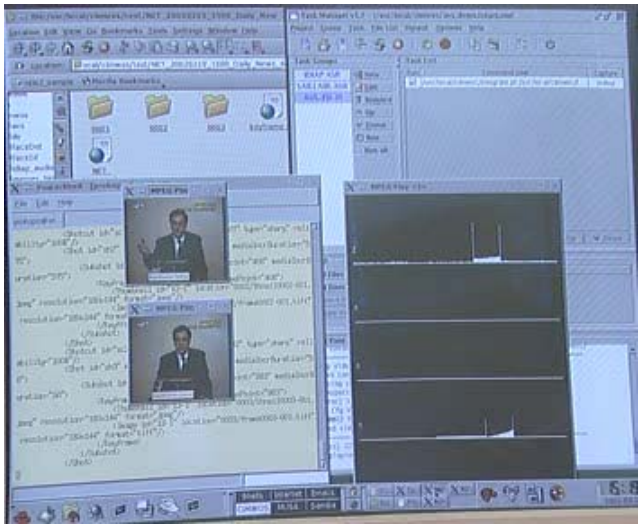


Video segmentation

A video sequence consists of many individual images, called frames, generally considered as the smallest unit to be concerned with, when segmenting a video. An 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 and identification

Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image, and if present, to return the image location and extent of the face. Face detection is a challenging task because of several factors influencing the appearance of the face in the image. These include identity, pose (frontal, half-profile, profile), presence or absence of facial features such as beards, moustaches and glasses, facial expression, occlusion and imaging conditions. Face detection has been, and still is, a very active research area within the computer vision community. With over 150 reported approaches to face detection, this research has broader implications for computer vision research on object recognition. It is one of the few attempts to recognise from images a class of objects for which there is a great deal of within-class variability. It is also one of the few classes of objects for which this variability has been captured using large training sets of images.

Recent descriptions of human face recognition by neuroscientists and psychologists show that face recognition is a dedicated process in human brain. This may have encouraged the view that artificial face recognition systems should also be face-specific. Automatic face recognition is a challenging task and has recently received significant attention. The rapidly expanding research in face processing is based on the recent development of technologies such as neural networks, wavelet analysis, and machine vision. Face recognition has a large potential for commercial application such as in authentication for banking, security system access, and advanced video surveillance.

In spite of expending research in face recognition, many problems remain unsolved, particularly in uncontrolled environments, because of lighting, facial expressions, background changes, and occlusion problems. One of the most important challenges in face recognition is to distinguish between intra-personal variation (in the appearance of a single individual due to a different expressions, lighting, etc.) and extra-personal variation (due to a difference in identity).

In CIMWOS, the face detection and face recognition modules associate faces occurring in video recordings with names. Both modules are based on “support vector machine” models trained on an extensive database of facial images with a large variation in pose and lighting conditions. Additionally, a semantic base has been constructed consisting of important persons that should be identified. During identification, images extracted from keyframes are compared to each model. At the decision stage, the scores resulting from the comparison are used either to identify a face or to reject it.

Text detection and recognition

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 colour, texture or shape), text usually conveys semantic information directly relevant to 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 colour 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 the background is necessary to fill the gap between images or video documents and the input of a standard OCR system.

In CIMWOS, the text detection and recognition module is based on a statistical framework using state-of-the-art machine learning tools and image processing methods. It consists of four modules:

1. Text detection aims at roughly and quickly finding blocks of image that may contain a single line of text characters.
2. Text verification, based on a Support Vector Machine model, tries to remove false alarms.
3. Text segmentation extracts pixels from text images belonging to characters with the assumption that they have the same colour/grey scale value. The method uses a Markov Random Field model and an Expectation Maximisation algorithm for optimisation.
4. Finally, all hypotheses produced by the segmentation algorithm are processed by the OCR engine. A string selection is made based on a confidence value, computed on the basis of character recognition reliability and a simple bigram language model.

Object recognition

The problem can be defined as the task of recognizing an object in a "recognition view," given knowledge accumulated from a set of previously seen "learning views." Until recently, visual object recognition was limited to planar (flat) objects, seen from an unknown viewpoint. The methods typically computed numerical geometric invariants from combinations of easily extractable image points and lines. These invariants were used as a model of the planar object for subsequent recognition. Some systems went beyond planar objects, but imposed limits on the possible viewpoints or on the nature of the object. Probably the only approach capable of dealing with general 3-D objects and viewpoints is the "Appearance-based" one, but it requires a very large amount of example views and has fundamental problems in dealing with cluttered recognition views and occlusion. A system capable of dealing with general 3-D objects, from general viewing directions is yet to be proposed.

Since 1998 a small number of works have emerged that seem to have the potential for reaching the general goal. These are all based on the concept of "region," defined as a small, closed area on the object's surface. In CIMWOS, the object surface is decomposed into a large number of regions automatically extracted from the images. 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 be gradually learned as new example views (or video streams) are acquired. The power of such a proposal consists fundamentally in two points: First, the regions themselves embed many small, local pieces of the object at the pixel level, and can reliably be put in correspondence along the example views. Even in case of occlusion or clutter in the recognition view, a subset of the object's regions will still be present. The second strong point is that the model captures the spatiotemporal 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 accumulate evidence about the identity of the object in the recognition view, even with only a small number of recognised regions.

Thanks to the good degree of viewpoint invariance of the regions, and to the strong model and learning approach developed, an object recognition system is obtained, to cope with 3-D objects of general shape, requiring only a limited number of learning views, that can recognise objects from a wide range of previously unseen viewpoints, in possibly cluttered, partially occluded, views.

Model

Similar views recovered automatically

Text processing

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, term recognition, story segmentation, and topic classification.

The task of the named entity detection module is to identify all named locations, persons and organisations as well as dates, percentage amounts and monetary amounts in the text produced by the speech recognition component. An initial finite-state preprocessor performs tokenisation and sentence boundary identification on the output of the speech recogniser. A part-of-speech tagger trained on a manually annotated corpus and a lexicon-based lemmatiser carry out morphological analysis and lemmatisation. A lookup module matches name lists and trigger-words against the text, and, eventually, a finite state parser recognises named entities on the basis of a pattern grammar.

The term recogniser identifies possible single or multi-word terms in the output of the speech processing system. 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.

Story detection and topic classification are performed using the same set of models, trained on an annotated corpus of stories and their associated topics. The basis of these technologies is a generative, mixture-based hidden Markov model that includes one state per topic and one state modelling general language, that is, words not specific to any topic. 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 a story the final state is reached. Detection 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 story boundaries, sections of text are classified according to a set of topic models.

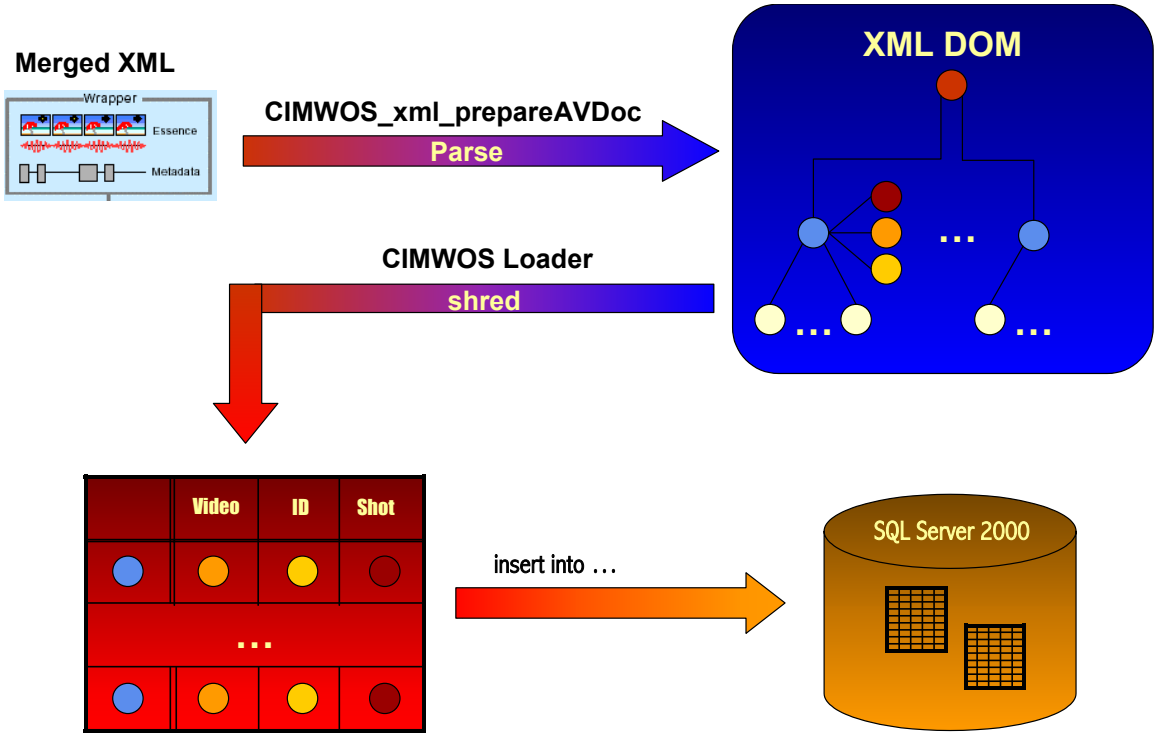
All technologies used are inherently language independent and of a statistical nature. The models were trained on several corpora collected and created within the CIMWOS project. The modelled inventory of topics is a flat, Reuters-derived structure containing about a dozen of main categories as well as several sub-categories.


```
<?xml version="1.0" ?>
- <results>
    <criteria Language="Greek" Text="πρόγνωση"
            Named_Entities="Παπαζάχος" Topics="γεωλογία, μετεωρολογία"
            Logical_Operator="OR" />
  - <passage score="100 %" number="1" media_title="News of 2002-05-24'
            passage_id="p44" startpoint_in_sec="529,05"
            duration_in_sec="15,76">
      <content>αναστάτωση αλλά και ανησυχίες προκαλεί στην κοινή
              γνώμη δημοσιοποίηση έκθεσης του Βασίλη του καθηγητή
              Βασίλη Παπαζάχου για μεσοπρόθεσμη πρόγνωση ισχυρός
              σεισμός σε τέσσερις περιοχές της Ελλάδας ο κύριος
              Παπαζάχος με δηλώσεις του ο υπερασπίστηκε την
              επιστημονική πέραση .</content>
      <speaker>male 22</speaker>
    - <media id="NET_20020524_1800_News" creation_date="2002-05-
              24T18:00+02:00" creator="NET" rights_owner="NET"
              availability_type="on demand delivery">
          <mediatitle>News of 2002-05-24</mediatitle>
          <source duration="1:74:52,48" />
      </media>
      <thumbnails number="1" />
  </passage>
```

Integration architecture

All processing output in the three modalities (audio, image and text) converges to a textual XML metadata annotation scheme following standard MPEG-7 descriptors. These annotations are further processed, merged, and loaded into the CIMWOS multimedia database. The merging component amalgamates the various XML annotations and creates a self-contained object compliant with the database. The resulting object can be dynamically transformed for interchanging semantic-based information into RDF and Topic Maps documents via XSL style sheets.

The CIMWOS retrieval engine is based on a weighted Boolean model equipped with intelligent indexing components. The basic retrieval unit is the passage, which has the role of a document in a traditional system. Therefore, 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. By linking shots to passages, each is assigned a broader set of features to be used for retrieval. Passages are represented as sets of features and retrieval is based on computed similarity in the feature space.



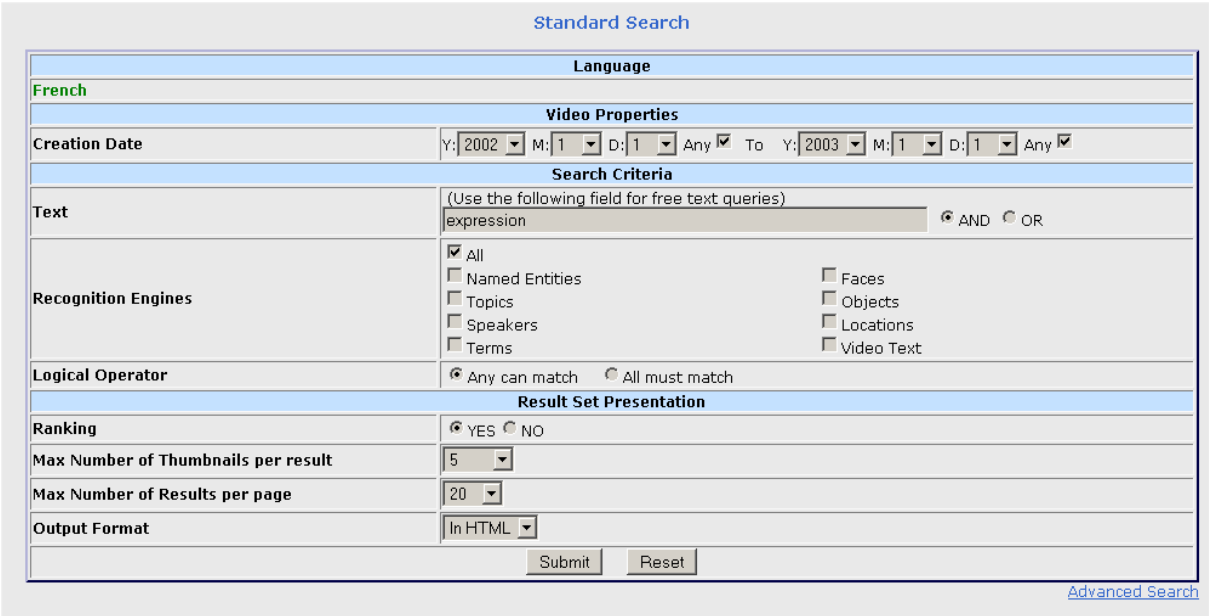
Search and retrieval

A video clip can take a long time to be transferred, e.g., from the digital video library to the user. In addition, it takes a long time to determine whether a clip meets one's needs. Returning half an hour of video when only one minute is relevant is much worse than returning a complete book when only one chapter is needed. Since the time to scan a video cannot be dramatically shorter than the real time of the video, it is important to give users only the material they need.

CIMWOS allows users to perform queries on the multimedia database by providing search criteria using any web browser. A "Simple query" mode acts as the standard query screen. This mode hides most of the complexity of the query mechanism and of the logical combination of a large amount of search criteria. An "Advanced query" mode is also available for expert users who want to monitor and keep track of the "search & query" process.

The retrieval system performs logical (Boolean) operations in order to process the user request. The operations are formed from queries and Boolean operators, such as "AND" and "OR." They represent a request to determine which documents contain the given set of keywords. In the retrieval procedure, a matching operation computes the similarity between the query and each passage. Finally, passages are ranked based on the result of the similarity computation.

The system returns a table of results. Each row of the table corresponds to one retrieved passage and includes the title of the video to which the passage belongs, the ID of the passage in the video, the duration of the passage in seconds, plus the predefined number of thumbnails (keyframes representative of the passage). A link is provided to the download site for this video, to launch the download and play procedure.



Advanced searches, browsing, detailed results

An advanced search allows the user to enter complex queries satisfying criteria specific to each type of processing. For example, different search keywords may be used for the speech transcription and the video text, allowing greater flexibility to search for a certain person (named in a legend on the screen) commenting on a certain topic (derived from the speech transcription). Additional fields are provided to enter restrictions on media information metadata, such as origin or duration.

The retrieval engine is then invoked with the complex set of criteria as combined flexibly with standard Boolean operators. The result set can be visualized by summarizing the relevant passages. While skimming a passage and its associated metadata, the end user can select a passage and view its associated metadata, the transcribed speech and results of the other processing components, as well as a representative sequence of thumbnails. Alternatively, the user may play the passage via intelligent video streaming.

In the case of databases with image and video data where subjectivity is a problem in indexing and identification, search by browsing becomes an important way of locating wanted items. Database browsing is closely related to the presentation of different types of abstraction of video that is stored in the database. The browsing facility of the CIMWOS system consists in a set of hyperlinks from each metadata and each attribute to the corresponding passages. When the user clicks on the number of a specific item, a new page presents full information of the corresponding video.

Video details contain the bibliographic metadata that describe the video source: creator and creation date, availability etc. Finally, the set of passage thumbnails is listed as part of the details page. Each thumbnail corresponds to a key-frame that is recognized and extracted by the video processing technology applied in CIMWOS.

<< Previous Video Segment Next Video Segment >>

Video Details

Source Video Info	
Creator	RTBF
Description	RTBF News Broadcast of 2003-08-05, 19h30
Creation Date	2003-08-05T19:30+00:30
Creation Location	Brussels
Rights Owner	RTBF
Total Video Duration	0:30:0
Availability Type	on demand delivery
Availability Start Date	2003-08-05
Availability End Date	2006-08-05
Video Segment Info	
Title	News of 2003-08-05 (English) Journal Televise 2003-08-05 (French)
Video Segment Duration	83,95 sec
Text	Boulangeries de sang de gré pour plus attachant un peu moins pour les chroniques de chaleur ici s' additionnent mieux vaut ne pas trop il pense . ... très chaud pour très froid ... le travailleur défend pas . Pour l'un ils changent d' autre le mille-feuille . ... le bon boulot ... comment la climatisées . Au DVD avaient leur va se mettre au soleil le royaume prend tout son futur . Quand même un des rares aujourd'hui apporter bonnet écharpe en volume nous nous avait déçu de chercher des vêtements croître le nombre d' enfants français ont trouvé souvent au nouveau chef de tee-shirts auteur d'un magasin de chercher de ce que va dire vous regarde . Tant au nord au sud de que marc garde 13 se disent ravis de leur métier ... que tous les conditions météo actuelles détails d'autres dit -il des surgelés il est un ancien boulanger qui a connu aussi la la promesse des métiers de véritables passions les trois ... .
Named Entities	DVD || soleil || marc
Topics	Amérique du Nord || Culture || Economie
Video Text	Marc Toncelli || Marc Toncelli || RTBF || Carry Responsable stock surgelés || Chalneux Paul || RTBF || 411 RTBF || RTBF || RTBF || Delvaux RTBF SImon || Liège RTBF || 0 Badart Reportage || RTBF || RTBF || RTBF
Format	Streaming Video (Requires MS Media Player)
Thumbnails	

Features and usability

Major features of the CIMWOS multimedia content management system include:

- Exploitation of multimodal, multimedia and multilingual information
- Indexing on image processing elements: faces, scene text, objects, video segments and keyframes, shots and subshots
- Indexing on speech processing elements: passages, speakers
- Indexing on text processing elements: words, terms, named entities, stories, topics
- Fully automated merging of multimodal metadata
- MPEG-7 compliant metadata
- XML-based indexing
- Alignment of indexes on time stamps
- Bibliographic XML indexing (media information)
- Advanced searching and browsing
- Weighted Boolean multimedia retrieval engine
- Combination of free query and metadata query
- Fast and efficient query execution
- Lexical analysis and stop lists
- Query normalisation for morphological variations
- Multilinguality: three European languages currently supported; architecture open for new languages to be included.
- Browsing schema

The CIMWOS integrated environment has been revised and updated based on the first validation report and on general feedback provided by the user groups participating in the project. Aspects such as user interface functionality and general user-friendliness, system reliability, and robustness in terms of accuracy and performance have been thoroughly tested and documented.

An expert external reviewer for a CIMWOS deliverable has commented as follows: "In the archive domain, the BBC has a standard for indexing the in-house production: 8 hours of a documenter for each hour of a produced program. With the help of the CIMWOS tools, I believe that the work could be reduced to 4 hours. I think this could be enough and sufficient to convince a General Manager having about 120.000 hours to digitise that this project is of strategic importance."

Project consortium

Role	Nº	Name	Short name	Country	Legal Status	Month in	Month out	Responsible for
CO	1	Institute for Language and Speech Processing	ILSP	EL	GOV	1	43	Project management, Greek text processing, search/retrieval, system integration, web site
CR	2	Matra Nortel Communications	MATRA	F	PRC/SA	1	1	Withdrawn
CR	3	Katholieke Universiteit Leuven	KULeuven	B	PNP	1	43	Shot cut detection, video segmentation, and face detection
CR	4	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur	LIMSI	F	GOV	1	12	Withdrawn
CR	5	Eidgenössische Technische Hochschule Zürich	ETHZ	CH	GOV	1	43	Object recognition, including template matching and region tracking
CR	6	Canal+ Belgique	Canal+	B	PRC/SA	13	43	User requirements and system validation
CR	7	Sail Labs Technology AG	SailLabs	A	PRC/AG	13	43	Speech recognition, text processing for French and English, dissemination and exploitation
CR	8	Institut Dalle Molle d'Intelligence Artificielle Perceptive	IDIAP	CH	PNP	13	43	Face recognition, text character detection and recognition, speech recognition

Project information

The project lasted 43 months, beginning in March 2000 and ending in September 2003. After a troubled start, due to partner acquisitions, a bankruptcy and two withdrawals, a new consortium was approved by the European Commission and continued the work specified in the original contract with minor modifications. The project completed all deliverables and often exceeded the original technological objectives by following the evolving state of the art throughout its duration.

CIMWOS was reviewed by the commission and approved to continue 3 times during its 3 active years: In September 2001 & 2002, and in April 2003. The project received very favourable comments each time. The final review took place in September 2003.

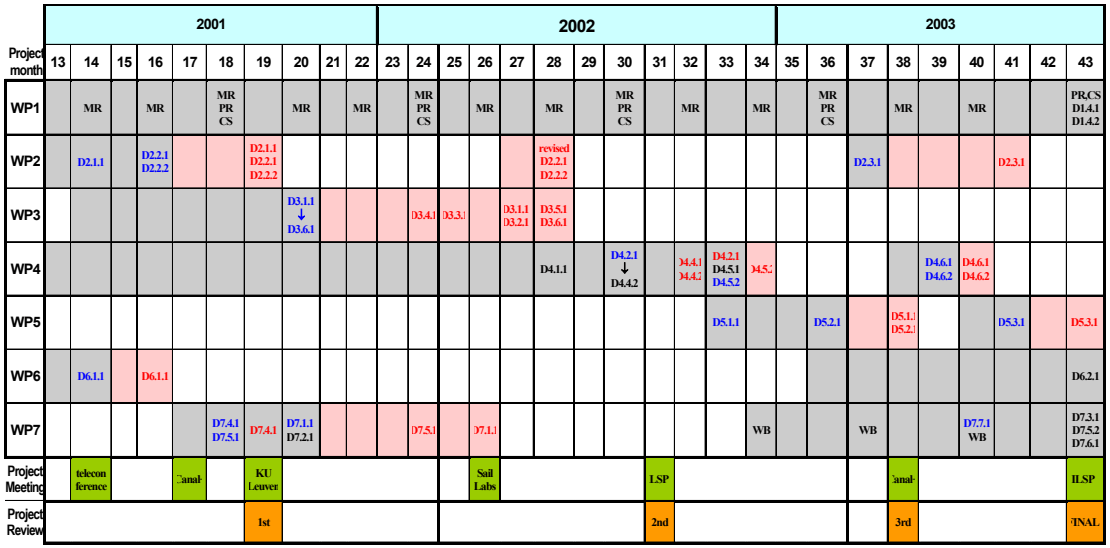
Seven meetings of the project coordination committee took place during the project lifetime, in addition to several teleconferences. Meetings were held in Athens, Brussels, Leuven, and Vienna. At the meetings, the project status was reviewed, deviations from the workplan were noted, and appropriate corrective action taken. In addition, design and implementation issues were thoroughly discussed many times, both among developers and between users and developers, before the final CIMWOS system as it exists today could possibly be realised.

The total budget of the project was 2.3 million euros, of which 1.0 million were contributed by the European Commission. Personnel costs (excluding overheads) accounted for approximately 60% of the total eligible project costs.

CIMWOS was coordinated by Dr. Athanassios Protopapas of the Institute for Language & Speech Processing, in Athens, Greece. For more information, he can be reached at ILSP, Artemidos 6 & Epidavrou, GR-151 25 Maroussi, Greece.

Tel: +30 210 6875409
Fax: +30 210 6854270
e-mail: protopap@ilsp.gr

The project web site is at www.xanthi.ilsp.gr/cimwos



Publications

V. Ferrari, T. Tuytelaars & L. Van Gool. Markerless Augmented Reality with a Real-time Affine Region Tracker. Proceedings of the IEEE and ACM International Symposium on Augmented Reality, New York, October 2001, pp. 87–96.

V. Ferrari, T. Tuytelaars & L. Van Gool Real-time Affine Region Tracking and Coplanar Grouping. Proceedings of the IEEE Computer Vision and Pattern Recognition, Hawaii, December 2001.

L. Van Gool, T. Tuytelaars, V. Ferrari, C. Strecha, J. Vanden Wyngaerd & M. Vergauwen. 3-D Modeling and Registration Under Wide Baseline Conditions, Proc. ISPRS Commission III, Vol. 34, Part 3A, Photogrammetric Computer Vision, Graz, September 2002, pp. 3–14.

L. Van Gool, D. Vandermeulen, G. A. Kalberer, T. Tuytelaars & A. Zalesny. Modeling Shape and Textures from Images: New Frontiers. 3-D Data Processing, Visualisation and Transmission 2002.

J.-M. Odobez & Datong Chen. Robust Video Text Segmentation and Recognition with Multiple Hypotheses. International Conference on Image Processing 2002.

F. Cardinaux & S. Marcel. Face Verification Using MLP and SVM. Neurosciences et Sciences de l'Ingenieur, France, November 2002.

S. Marcel & S. Bengio. Improving Face Verification using Skin Color Information. 16th International Conference on Pattern Recognition, 2002

F. Cardinaux, C. Sanderson & S. Marcel. Comparison of MLP and GMM Classifiers For Face Verification on XM2VTS. AVBPA 2003.

N. Poh, S. Marcel & S. Bengio. Improving Face Authentication Using Virtual Samples. ICASSP 2003.

D. Chen, J-M. Odobez, & J-P. Thiran. A Localization/ Verification Scheme for Finding Text in Images and Videos Based on Contrast Independent Features and Machine Learning Methods. To appear in Image Communication Journal.

V. Ferrari, T. Tuytelaars, & L. Van Gool. Wide-baseline muliple-view correspondences. IEEE Computer Vision and Pattern Recognition, Madison, USA, June 2003.

V. Ferrari, T. Tuytelaars, & L. Van Gool. Exploring Images for Object Recognition, submitted to the European Conference on Computer Vision.

H. Shao, T. Svoboda, V. Ferrari, T. Tuytelaars, & L. Van Gool. Fast indexing for image retrieval based on local appearance with re-ranking. International Conference on Image Processing, Sept. 2003.

D. Chen, J-M. Odobez, & H. Bourlard. Text Detection and Recognition in Images and Videos. To appear in the Pattern Recognition Journal.

D. Chen & J-M. Odobez. Monte Carlo Video Text Segmentation. ICIP 2003.

I. McCowan, J. Ajmera & D. Moore. An Online System for Automatic Annotation of Audio Documents. IDIAP Research Report 03-39. Martigny, Switzerland, 2003.

H. Papageorgiou & A. Protopapas. CIMWOS: A multimedia, multimodal and multilingual indexing and retrieval system. In Ebroul Izquierdo (ed.) Digital Media Processing for Multimedia Interactive Services (pp. 563-568); World Scientific Publishing Co., 2003

H. Papageorgiou, A. Protopapas, & T. Netousek. Retrieving video segments based on combined text, speech and image processing. 2003 Broadcast Engineering Conference Proceedings, pp. 177–182.

J. Riedler & S. Katsikas. Automatic Speech Recognition of Modern Greek Broadcast News. Submitted to Speech Communication.