The CIMWOS Multimedia Indexing System

Harris Papageorgiou and Athanassios Protopapas

Institute for Language and Speech Processing
Artemidos 6 and Epidavrou, Athens 15125, Greece
{xaris,protopap}@ilsp.gr

Abstract. We describe a multimedia, multilingual and multimodal research system (CIMWOS) supporting content-based indexing, archiving, retrieval and on-demand delivery of audiovisual content. CIMWOS (Combined IMage and WOrd Spotting) incorporates an extensive set of multimedia technologies by seamless integration of three major components – speech, text 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 documents via XSL stylesheets. The CIMWOS Retrieval Engine is based on a weighted boolean model with intelligent indexing components. A 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 includes sports, broadcast news and documentaries in three languages.

1 Introduction

The advent of multimedia databases and the popularity of digital video as an archival medium pose many technical challenges and have profound implications for the underlying model of information access. Digital media assets are proliferating and most organizations, large broadcasters as well as SMEs are building networks and technology to exploit them. Traditional broadcasters, publishers and Internet content providers are migrating into increasingly similar roles as multimedia content providers. Digital technology today allows the user to manipulate or interact with content in ways not possible in the past. The combination of PCs and networks allows the individual to create, edit, transmit, share, aggregate, personalize and interact with multimedia content in increasingly flexible ways. The same technology allows content to be carried across different platforms. In fact, much of the information that reaches the user nowadays is in digital form: digital radio, music CDs, MP3 files, digital satellite and digital terrestrial TV, personal digital pictures and videos and, last but not least, digital information accessed through the Web. This information is heterogeneous, multimedia and, increasingly, multi-lingual in nature.

The development of methods and tools for content-based organization and filtering of this large amount of multimedia information reaching the user through many and different channels is a key issue for its effective consumption and enjoyment. There are several projects aiming at developing advanced technologies and systems to tackle the problems encountered in multimedia archiving and indexing [1], [2], [3].

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 619–627, 2004.

The approach taken in CIMWOS was to design, develop and test an extensive set of multimedia technologies by seamless integration of three major components – speech, text and image processing – producing a rich collection of XML metadata annotations and allowing the user to store, categorize and retrieve multimedia and multilingual digital content across different sources (TV, radio, music, Web).

This paper is organized as follows: in the next three sections we focus on technologies specific to Speech, Text, and Image respectively. These technologies incorporate efficient algorithms for processing and analyzing relevant portions from various digital media and thus generating high-level semantic descriptors in the metadata space. CIMWOS architecture for the integration of all results of processing is presented in the following section. Evaluation results are reported in the next section and finally future work is drawn in the last section.

2 Speech Processing Component

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. The segmentation of the continuous stream into homogeneous sections (speaker and/or acoustic/background conditions) also poses serious problems. Successful segmentation however, forms the basis for further adaptation and processing steps. Consequently, adaptation to the varied acoustic properties of the signal or to a particular speaker, and enhancements of the segmentation process, are generally acknowledged to be key areas for research to render indexing systems usable for actual deployment. This is reflected by the effort and the number of projects dedicated to advance the state-of-the-art in these areas.

In CIMWOS, all of these tasks are being taken care by the Speech Processing Component (SPC). The SPC comprises 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 [4], trained in a wide variety of noise conditions encountered in the broadcast news domain.

3 Text Processing Component

After processing the audio input, text-processing tools operate on the textual stream produced by the SPC and perform the following tasks: Named Entity Detection (NED), Term Extraction (TE), Story Detection (SD) and Topic Classification (TC).

3.1 Named Entity Detection (NED) and Term Extraction (TE)

The task of the Named Entity Detection (NED) module is to identify all named locations, persons and organizations in the transcriptions produced by the ASR component. CIMWOS uses a series of basic language technology building blocks, which are modular and combined in a pipeline [5]. An initial finite state preprocessor performs tokenization on the output of the speech recognizer. A part-of-speech tagger trained on a manually annotated corpus and a lemmatizer carry out morphological analysis and 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 NED module includes populating the gazetteer lists and semi-automatically extracting the pattern rules. A corpus of 100.000 words of gold transcriptions of broadcast news per language (English, Greek) already tagged with the NE classes was used to guide system training and development.

The term extraction (TE) module involves the identification of single or multi-word indicative keywords (index terms) in the output of the ASR. Systems for automatic term extraction using both linguistic and statistical modeling are reported in the literature [6]. The term extractor in CIMWOS follows the same architecture: linguistic processing is performed through an augmented term grammar, the results of which are statistically filtered using frequency-based scores.

NED and TE modules were tested on pre-selected sequences of Greek broadcasts [10]. The NED module obtained a 79-80% precision and a 52-53% recall for locations and persons. Lower recall is due to missing proper names in the vocabulary of the ASR engine as also the diversity of domains found in broadcasts. The TE module automatically annotated the same data, scoring a 60% recall and a 35% precision.

3.2 Story Detection (SD) and Topic Classification (TC)

The basis of the Story Detection (SD) and Topic Classification (TC) modules is a generative *mixture-based Hidden Markov Model* (HMM). The HMM includes one state per topic and one state modeling 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 the story, a final state is reached. SD is performed running the resulting models on a fixed-size sliding window, noting changes 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 thus located, TC classifies the sections of text according to the set of topic models (and a general language model). The modeled inventory of topics is a flat, Reuters-derived structure containing about a dozen of main categories and 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 shallow enough to provide reasonable amounts of training data per topic but still fine-grained enough to allow for flexibility and detail in queries.

4 Image Processing Component

The Image Processing Component (IPC) consists of modules responsible for video segmentation and keyframe extraction; detection and identification of faces at any location, scale, and orientation; recognition of objects given knowledge accumulated from a set of previously seen "learning views"; and video text detection and recognition. A brief description of these modules is given in the following subsections.

4.1 Face Detection and Identification (FD/FI)

The FD and FI modules spot faces in keyframes and associate them with names. Given a keyframe extracted from the video by the AVS module, the FD module will attempt 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. In spite of expanding research in the field, FD remains a very 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. In face recognition and identication, intra-personal variation (in the appearance of a single individual due to different expressions, lighting, etc.) and extra-personal variation (variations in appearance between persons) are of particular interest.

In CIMWOS, both FD and FI modules [7] are based on Support Vector Machine models. The FD module was trained on an extensive database of facial images with a large variation in pose and lighting conditions. Additionally, a semantic base of important persons was compiled for FI training. During the FI stage, faces detected by the FD module in the keyframes are compared to the model corresponding to each face in the semantic base. Scores resulting from the comparison guide the module to identify a candidate face or classify it as "unknown."

4.2 Object Recognition (OR)

Object Recognition (OR) is used to spot and track pre-defined "objects" of interest. The objects can be scenes, logos, designs, or any user-selected image parts of importance, manually delineated on a set of "example views," which are used by the system to create object models. In CIMWOS, the object's surface is decomposed in a large number of regions (small, closed areas on the object's surface) automatically extracted from the keyframes. The spatial and temporal relationships of these regions are acquired from several example views. These characteristics are incorporated in a model, which can thus be gradually augmented as more object examples are assimilated [8].

This methodology has two fundamental advantages: first, the regions themselves embed many small, local pieces of the object appearance at the pixel level. Thus, even in the case of occlusion or clutter, they can reliably be associated with training examples, since a subset of the object's regions will still be present. The second strong point is that the model 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 accumulate evidence about the identity of the object in the recognition view, even in cases where only a small amount of recognized regions is found (e.g.: strong occlusion, difficult illumination conditions, which might make many individual regions hard to spot).

Thanks to the good degree of viewpoint invariance of the regions, and to the strong model and learning approach developed, the OR module can cope with 3-D objects of general shape, requiring only a limited number of learning views. This way objects can be recognized in a wide range of previously unseen viewpoints, in possibly cluttered, partially occluded, views.

4.3 Video Text Detection and Recognition (TDR)

Text in a video stream may appear in captions produced by the broadcaster, or in labels, logos etc. The goal of the TDR module is to efficiently detect and recognize these textual elements by integrating advanced Optical Character Recognition (OCR) technologies. Since text often conveys semantic information directly relevant to the content of the video (such as a politician's name or an event's date and location), TDR is recognized as a key component in an image/video annotation and retrieval system. However, text characters in video streams usually appear against complex backgrounds and may be of low resolution, and/or of any colour or greyscale value. The direct application of conventional OCR technology has been shown to lead to poor recognition rate. Better results are obtained if efficient location and segmentation of text characters occurs before the OCR stage.

The CIMWOS TDR module is based on a statistical framework using state-of-the-art machine learning tools and image processing methods [9]. Processing consists of four stages: Text detection aims at roughly and quickly finding blocks of image that may contain a single line of text characters. False alarms are removed during the text verification stage, on the basis of a Support Vector Machine model. Text segmentation uses a Markov Random Field model and an Expectation Maximization algorithm to extract pixels from text images belonging to characters, with the assumption that they have the same colour/grey scale value. During the final processing stage, 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.

5 Integration Architecture

The critical aspect of indexing the video segment is the integration of image and language processing. Each scene is characterized by the metadata that appear in it. 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 merging Component of CIMWOS is responsible for the amalgamation of the 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 into RDF (http://www.w3.org/RDF/) documents, for interchanging semantic-based information, via XSL stylesheets (http://www.w3.org/Style/XSL).

The CIMWOS Integration architecture (Fig 1) follows a N-tier scenario by integrating a data services layer (storage & retrieval of metadata), a business services layer incorporating all remote multimedia processors (audio, video and text intelligent engines), and a user services layer which basically includes the user interface (UI) and web access forms.

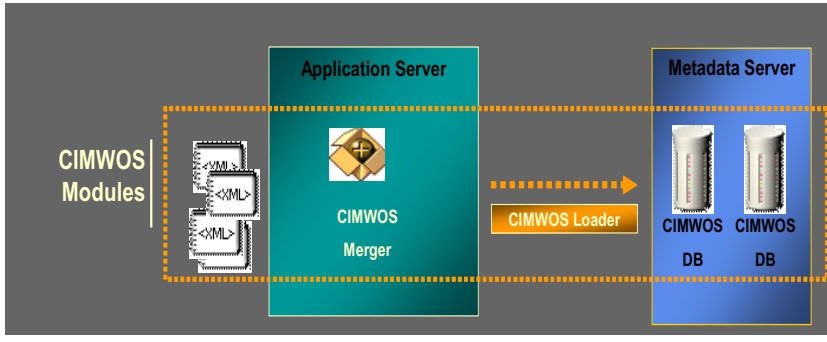

Fig. 1. CIMWOS Architecture

CIMWOS retrieval engine is based on a weighted Boolean model. Each video segment is represented by its XML metadata annotation object. Boolean logic is effectively combined with a metadata-weighting scheme for the similarity measure which is a function of three factors: metadata-level weights based on the calculated overall precision of each multimedia processor, value-level statistical confidence measures produced by the different engines and tf*idf scores for all textual elements (spoken words). A more detailed description of the retrieval engine can be found in [10].

6 Evaluation

In this section, we focus on the assessment of the Multimedia Information Retrieval (MIR) CIMWOS system response to user queries. The evaluation of each contributing module is reported elsewhere [10]. The accuracy of the integrated environment's retrieval capabilities were tested on Greek news broadcasts, on the basis of groundtruth annotations created by users of different expertise.

6.1 Evaluation Materials and Method

Evaluation and validation were divided in two phases. During the first phase, the overall success in retrieval of passages relevant to a particular topic was assessed on pre-selected sequences of Greek broadcasts. During the second phase of the validation the retrieval task was repeated, this time on new video material, for which boundaries of relevant stories for each topic had been previously identified by human annotators.

A group of three individuals (2 men, 1 woman) was set up to carry our validation from an end-user viewpoint. They had different extensive backgrounds in working with multimedia objects: an archivist of audiovisual material for the Greek state television, a postgraduate student in journalism and media studies, and a historian specialized in the creation and management of historical archives.

We used 35.5 hours of digital video during the two validation phases. The material consisted of Greek news broadcasts by state and private TV networks between March 2002 and July 2003. Each video file corresponded to one news broadcast. For the first validation phase we used 15 videos (henceforth, Collection A) of approximately 18 hours total duration, captured in BETA SP and transcribed in MPEG-2. For the second phase we used 15 broadcasts (henceforth, Collection B) amounting to approximately 17 hours of video, captured via standard PC TV cards in MPEG-2. Bibliographic data (creator, duration and format, etc.) were recorded in the project's media description format. The fact that the video collection spanned a relatively long time period ensured the diversity of the news stories presented in the broadcasts.

Gold annotations of each collection were created by the user group, using XML-aware editors for the compilation of groundtruth data. The CIMWOS query interface offers different views of the results, and each user was responsible for storing results in reusable XML files containing information about the criteria used in each query. A user familiar with the test collection was responsible for the generation of a list of interesting topics. The user then located manually all relevant sequences, allowing the association of start and end timestamps with each topic.

During search, users generated queries for each topic of the list based on their own judgment, using combinations of Terms, Named Entities and/or ASR Text. Users formed 5 queries, on average, for each topic. Using the HTML interface, they could browse the results to assess their overall satisfaction with the results.

Users found the returned passages to be short and fragmented in the results from queries on Collection A videos. This was attributed to the fact that passage identification was based on automatic segmentation and clustering by the speech recognition module, based on speaker turns. For Collection B testing, a different approach was taken to produce more intuitive passage segmentation. Manual identification of relevant segments was again undertaken, but this time segmentation was based on the *stories* contained in each broadcast. The story boundaries were then aligned to the ASR transcriptions, thus avoiding the temporally short passages observed phase A.

6.2 Results

We collected each version for each user's query as an XML file, and tested against gold data. Results are shown in Table 1.

Further testing included filtering out results that scored less than 60% in the CIMWOS DB ranking system. Although a decrease in the system's recall was observed in the case of Collection B, this filter significantly increased precision in both validation phases.

The MIR validation phase confirmed the hypothesis that not all metadata annotations are equally important in terms of retrieval accuracy and users' satisfaction. The experiments showed that accurate TDR and OR combined with a state-of-the-art ASR engine (10-20% WER in BNs) can adequately support most retrieval tasks specifically in case the search unit is the story and not the passage (speaker turns). Moreover, FD/FI information should be combined with ASR/NED and SID (speaker identification) results in order to increase the accuracy of the named persons.

Table 1. Retrieval results on Greek video collections

	Precision	Recall	F-measure
Collection A	34.75	57.75	43.39
Collection A + 60% Filter	45.78	53.52	49.35
Collection B	44.78	50.24	47.36
Collection B + 60% Filter	64.96	37.07	47.20

7 Future Work

The three major components – speech, text and image – of the multimodal indexing subsystem incorporated in the CIMWOS integrated platform produce a rich set of metadata indices following MPEG-7 descriptors, allowing for flexibility in retrieval tasks. Our future work focuses on semantically enriching the contents of multimedia documents with topic, entity and fact information relevant to user profiles; developing suitable cross-language cross-media representations; and building classification and summarization capabilities incorporating cross-language functionality (cross-language information retrieval, categorization and machine translation of indicative summaries) based on statistical machine translation technology.

Acknowledgments. This work was supported in part by shared-cost research and technological development contract IST-1999-12203 with the European Commission (project CIMWOS; see www.xanthi.ilsp.gr/cimwos/).

CIMWOS subsystems and components developed at the Institute for Language & Speech Processing (Greece), Katholieke Universiteit Leuven (Belgium), Eidgenossiche Technische Hochschule Zurich (Switzerland), Sail Labs Technology AG (Austria), and Institut Dalle Molle d'Intelligence Artificielle Perceptive (Switzerland). User requirements compiled by Canal+ Belgique (Belgium).

References

1. 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
2. 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.
3. 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
4. Kubala, F., Davenport, J., Jin, H., Liu, D., Leek, T., Matsoukas, S., Miller, D., Nguyen, L., Richardson, F., Scwhartz, R. & Makhoul, J. (1998). The 1997 BBN BYBLOS System applied to Broadcast News Transcription. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.* Lansdowne VA.
5. Demiros, I., Boutsis, S., Giouli, V., Liakata, M., Papageorgiou, H. & Piperidis, S. (2000) Named Entity Recognition in Greek Texts. In *Proceedings of Second International Conference on Language Resources and Evaluation-LREC2000* (pp.1223-1228). Athens, Greece.
6. Jacquemin, C. & Bourigault, D. (2003). Term Extraction and Automatic Indexing. In Mitkov R., (Ed.), *Handbook of Computational Linguistics*, (pp. 599-615). Oxford University Press, Oxford.
7. Cardinaux F. & Marcel S. (2002). Face Verification Using MLP and SVM. In Neurosciences et Sciences de l'Ingenieur, France.
8. Ferrari, V., Tuytelaars, T., & Van Gool, L. (2003). Wide-baseline muliple-view correspondences. In *Proc. of IEEE Computer Vision and Pattern Recognition*. Madison, USA.
9. Odobez, J. M., & Chen, D. (2002) Robust Video Text Segmentation and Recognition with Multiple Hypotheses. In *Proc. of the International Conference on Image Processing.*
10. Papageorgiou H., Prokopidis, P., Demiros I., Hatzigeorgiou N. & G. Carayanis. (2004). CIMWOS: A Multimedia retrieval system based on combined text, speech and Image processing. In *Proceedings of the RIAO Conference (RIAO-2004)*, Avignon, France