Ebroul Izquierdo, editor. ***Digital Media Processing for Multimedia Interactive Services***.
Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services.
World Scientific Publishing Co., 2003.

CIMWOS: A MULTIMEDIA, MULTIMODAL AND MULTILINGUAL INDEXING AND RETRIEVAL SYSTEM

H. PAPAGEORGIOU AND A. PROTOPAPAS

Institute for Language & Speech Processing
Artemidos 6 & Epidavrou,
GR-151 25 Maroussi, Greece
E-mail: {xaris, protopap}@ilsp.gr

CIMWOS is a multimedia, multimodal and multilingual system supporting content-based indexing, archiving, retrieval, and on-demand delivery of audiovisual content. 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 audio processing operations employ 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. The retrieval engine is based on a weighted boolean model with intelligent indexing components. An ergonomic and user-friendly web-based interface allows the user to efficiently retrieve video segments by a combination of media description, content metadata and natural language text. The system is currently under evaluation.

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 models of information access. There are several projects developing advanced technologies for multimedia archiving and indexing but the available tools and applications to describe, organize, and manage video data remain limited [1–4]. CIMWOS[a] (Combined IMage and WOrd Spotting) incorporates an extensive set of multimedia technologies, integrating three major subsystems – text, speech, and image processing – and

[a] Project web site: http://www.xanthi.ilsp.gr/cimwos/

producing a rich collection of XML metadata annotations. The annotations are merged and loaded into a multimedia database, where they can be searched and retrieved via weighted boolean searches entered in a web-based interface.

2. Processing subsystems

2.1. *Speech Processing*

Broadcast news exhibit a wide variety of audio characteristics, including 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 when, what topic a segment is about, or which organisations are mentioned, are all challenging problems. Adverse background conditions can lead to significant degradation in performance. Consequently, adaptation to the varied acoustic properties of the signal or to a particular speaker, and enhancements to the segmentation process, are key areas for research and improvement to render indexing systems usable. This is reflected in the effort dedicated to advance the state-of-the-art in these areas.

In CIMWOS, these tasks are handled by the speech processing subsystem, which comprises rejection of music [5], speaker change detection (SCD), automatic speech recognition (ASR), speaker identification (SID) and speaker clustering (SC). The ASR engine is a real-time, large vocabulary, speaker-independent, gender-independent, continuous speech recogniser [6] trained in a wide variety of noise conditions encountered in the broadcast news domain.

2.2. *Text Processing*

Text processing tools, operating on the text stream produced by the speech processing subsystem, perform detection, identification, and classification tasks.

2.2.1. Named Entity Detection (NED)

The NED module identifies all named locations, persons and organisations, as well as dates, percentages and monetary amounts in the speech transcriptions. CIMWOS uses a series of basic language technology building blocks, which are modular and combined in a pipeline [7]. 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 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. 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 recogniser.

2.2.2. Term Recognition (TR)

The term recogniser identifies possible single or multi-word terms, using both linguistic and statistical modelling. Linguistic processing is based on an augmented term grammar, the results of which are statistically filtered using frequency-based scores. Preliminary testing shows the method capable of locating 62% of technical terminology.

2.2.3. Story Detection and Topic Classification (SD/TC)

Story detection (SD) and topic classification (TC) use a set of models trained on an annotated corpus of stories and their associated topics. The basis of SD and TC is a generative, mixture-based HMM including one state per topic and one state modelling general language, i.e., words not specific to any 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. In SD, a sliding window of fixed size is used to note the change in topic-specific words, resulting in a set of 'stable regions' in which topics change only slightly or not at all. TC then classifies the text sections according to the set of topic models.

The modelled inventory of topics is a flat, Reuters-derived structure, containing a few main categories and several sub-categories, a structure shallow enough to provide reasonable amounts of training data per topic but still fine-grained enough to allow for flexibility and detail in queries. All technologies used are inherently language independent and of a statistical nature.

2.3. *Image Processing*

2.3.1. Video Segmentation (AVS)

A video sequence consists of many individual images (frames); an uninterrupted video stream generated by one camera is called a shot. A shot cut is the point at

which shots change within a video sequence. Video segmentation partitions the raw material into shots by measuring the differences between consecutive frames and applying adaptive thresholding on motion and texture cues. For each shot a few representative frames are selected, referred to as keyframes. Keyframes contain most of the static information present in a shot, so face recognition and object identification can focus on keyframes only.

2.3.2. Face Detection (FD) and Identification (FI)

The FD and FI modules associate faces occurring in video recordings with names. Significant challenges are posed because of differences in illumination, facial expressions, background, and occlusion. A central problem in face recognition is to distinguish between variations in appearance of the same person due to such differences from variations between different persons. FD and FI modules are trained on a database of facial images with a large variation in pose and lighting conditions [8]. Additionally, a semantic base has been constructed consisting of important persons that the FI module should identify.

2.3.3. Object Recognition (OR)

Object Recognition is used to spot and track pre-defined objects of interest. Surfaces are decomposed into regions that are automatically extracted from the images [9,10] from several viewpoints (or frames) and are incorporated in a model. Models 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 recognised regions is found such in cases of strong occlusion, difficult illumination conditions etc.

2.3.4. Video Text Detection and Recognition (TDR)

Text detection and recognition in images and video integrates optical character recognition (OCR) with text-based search. TDR is 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 about the video contents, such as a player's or speaker's name, location and date of an event, etc. The CIMWOS TDR module is based on a statistical framework [11,12], and performs text detection, to quickly find relevant blocks of images; text verification, to remove false alarms; text segmentation, to extract pixels belonging to characters; and finally OCR, to recognise the characters.

3. INTEGRATION ARCHITECTURE

All processing in the three modalities (audio, image and text) converges to a textual XML metadata annotation scheme following MPEG-7 descriptors [13]. 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 database is a large collection of broadcast news and documentaries in three languages (English, Greek, and French), though more languages can be added.

4. RETRIEVAL AND PRESENTATION

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.

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. 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. In the retrieval procedure, passages are first filtered on any selected advanced features to reduce the search space of the free queries. Next, a boolean-based matching operation computes the similarity between the query and each passage. Finally, passages are ranked based on the result of the similarity computation. 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, a summary containing the first words of the transcribed speech and a sequence of thumbnails, or (s)he can play the passage via intelligent streaming.

Acknowledgments

Work supported by cost reimbursement contract number IST-1999-12203 for research and technological development projects from the Commission of the European Communities, Directorate-General Information Society.
CIMWOS subsystems and components developed at the Institute for Language & Speech Processing (Greece), Katholieke Universiteit Leuven (Belgium), Eidgenössische Technische Hochschule Zürich (Switzerland), Sail Labs Technology AG (Austria), and Institut Dalle Molle d'Intelligence Artificielle Perceptive (Switzerland). User requirements by Canal+ Belgique (Belgium).

References

1. H. Wactlar, A. Olligschlaeger, A. Hauptmann and M. Christel. *Comm. ACM*, **43**, 42 (2000).
2. M.R. Lyu , E. Yau and S. Sze. *Proc. 2nd ACM/IEEE-CS Joint Conf. Digital Libraries*, 145 (2002).
3. A. Sankar, R.R. Gadde and F. Weng. *Proc. DARPA Broadcast News Workshop*, 281 (1999).
4. N. Dimitrova, H.J. Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhor. *IEEE MULTIMEDIA*, **9**, 42 (2002).
5. J. Ajmera, I. McCowan and H. Bourlard. *Proc. ICASSP* (2002).
6. F. Kubala, J. Davenport, et al. *Proc. DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne VA* (1998).
7. I. Demiros, S. Boutsis, V. Giouli, M. Liakata, H. Papageorgiou and S. Piperidis. *Proc. 2nd Int. Conf. Language Resources and Evaluation*, Athens, Greece, 1223 (2000).
8. F. Cardinaux and S. Marcel. *XI Journees NeuroSciences et Sciences pour l'Ingenieur* (2002).
9. T. Tuytelaars, A. Zaatri, L. Van Gool and H. Van Brussel. *IEEE Conf. Robotics and Automation*, 3707 (2000).
10. V. Ferrari, T. Tuytelaars, L. Van Gool. *IEEE Conf. Computer Vision Pattern Recognition*, 226 (2001).
11. D. Chen. H. Bourlard and J.P. Thiran. *Proc. Int. Conf. Computer Vision Pattern Recognition*, 621 (2001).
12. J.M. Odobez and D. Chen. *Proc. ICIP* (2002).
13. H. Papageorgiou, P. Prokopidis, I. Demiros, G. Giouli, A. Constantinidis, and S. Piperidis. *Proc. 3rd Language Resources and Evaluation Conf.*, Las Palmas, 1723 (2002).