11

Multimedia Indexing and Retrieval Using Natural Language, Speech and Image Processing Methods

Harris Papageorgiou, Prokopis Prokopidis, Athanassios Protopapas and George Carayannis

11.1 Introduction

Throughout the chapter, we provide details on implementation issues of practical systems for efficient multimedia retrieval. Moreover, we exemplify algorithms and technologies by referring to practices and results of an EC-funded project called Combined IMage and WOrd Spotting (CIMWOS) [1], developed with the hope that it would be a powerful tool in facilitating common procedures for intelligent indexing and retrieval of audiovisual material. CIMWOS used a multifaceted approach for the location of important segments within multimedia material, employing state-of-the-art algorithms for text, speech and image processing in promoting reuse of audiovisual resources and reducing budgets of new productions.

In the following three sections, we focus on technologies specific to speech, text and image, respectively. These technologies incorporate efficient algorithms for processing and analysing relevant portions from various digital media and thus generating high-level semantic descriptors in the metadata space. After proposing an architecture for the integration of all results of processing, we present indicative evaluation results in the context of CIMWOS. Relevant Content-based Information Retrieval (CBIR) research prototypes and commercial systems are briefly presented in the Section 11.7.

Multimedia Content and the Semantic Web Edited by Giorgos Stamou and Stefanos Kollias
 ISBN: 0-470-85753-6.

11.2 Audio Content Analysis

Audio content analysis for multimedia indexing and retrieval refers to a number of audio features gathered from acoustic signals. These audio features combined with visual descriptors have effectively been exploited in scene content analysis, video segmentation, classification and summarization [2]. Audio features are extracted in the short-term frame level and the long-term clip level (usually called window). A frame is defined as a group of adjacent samples lasting about 10–40 ms, within which the audio signal is assumed to be stationary. Frame-level features are:

- *Time-domain features*: volume, zero crossing rate, pitch.
- *Frequency-domain features* (based on the fast fourier transform (FFT) of the samples in a frame): spectrum of an audio frame, the ratio of the energy in a frequency subband to the total energy (ERSB), the mel-frequency cepstrum coefficients (MFCC) widely used for speech recognition and speaker recognition.

Clip-level features capture the temporal variation of frame-level features on a longer time scale and include volume-based features like the VSTD-standard deviation of the volume over a clip, the ZSTD proposed by Liu et al. [3] and pitch-based features like the SPR, which is the percentage of frames in a clip that have similar pitch as the previous frames.

Audio tracks are further segmented and classified into a number of classes such as speech, silence, noise, music and environmental sound (recognition of shots, cries, explosions, door slams etc.). State-of-the-art continuous speech recognizers perform on the speech segments, generating speech transcriptions. Producing a transcript of what is being said, and determining who is speaking and for how long are all challenging problems, mainly due to the continuous nature of an audio stream: segments of diverse acoustic and linguistic nature exhibit a variety of problematic acoustic conditions, such as spontaneous speech (as opposed to read or planned speech), limited bandwidth (e.g. telephone interviews), and speech in the presence of noise, music and/or background speakers. Such adverse background conditions lead to significant degradation in the performance of speech recognition systems if appropriate countermeasures are not taken. Likewise, segmentation of the continuous speech stream into homogeneous sections (with respect to acoustic/background conditions, speaker and/or topic) poses serious problems. Successful segmentation, however, forms the basis for further adaptation and processing steps. Adaptations to the varied acoustic properties of the signal or to a particular speaker, and enhancements to the segmentation process, are generally acknowledged [4] as key research areas that will result in rendering indexing systems usable for actual deployment. This is reflected in the amount of effort and the number of projects dedicated to the advancement of the current state-of-the-art in these areas [5].

A speech processing subsystem (SPS) usually comprises a set of technology components responsible for a particular task, arranged in a pipeline. These tasks typically include speaker change detection (SCD), automatic speech recognition (ASR), speaker identification (SID) and speaker clustering (SC). To save processing time, SC and SID can be run in parallel to ASR. Input to the overall system is the audio stream, while the final output is a set of audio descriptors containing speech transcriptions and identified speakers. A typical SPS architecture is depicted in Figure 11.1, while each specific task is presented in detail in the following subsections.

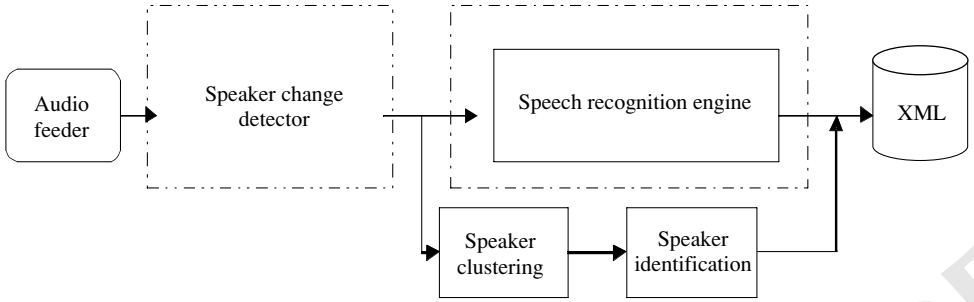

Figure 11.1 Speech processing subsystem

11.2.1 Speaker Change Detection

SCD aims at extracting only the speech portions of the incoming audio stream, and at partitioning those sections into homogeneous sections of speech. The technology used for SCD involves initially running the speech recognizer with a phone-level model (PLD), which is a set of broad phonetic classes of speech sounds like vowels, nasals, obstruents, as well as some non-speech sounds like music, laughter etc.

Using the information produced by the PLD, the SCD module determines, during runtime, likely boundaries of speaker turns based on non-speech portions (e.g. silence, music, background noise) in the signal. Parameters used in this decision are usually empirically determined, as they differ from language to language and even from domain to domain.

11.2.2 Speaker Clustering and Identification

The SC module is responsible for grouping predetermined chunks of audio (the SCD output described above) into a number of clusters. Resulting clusters ideally represent utterances produced by a single speaker. SID uses a predefined set of target models to identify the speaker of a segment or, if identification is not possible, to determine the speaker's gender. The set of target speakers typically comprises anchor speakers of TV stations as well as persons of public interest such as politicians. Current SID technology employs Gaussian mixture models (GMM), trained on corpora of speaker-annotated transcriptions. On the other hand, SC does not use any trained models, but rather a variant of the generalized likelihood ratio criterion at runtime. Evaluation of the CIMWOS SID module on predefined test sets produced 97% precision and 97% recall figures for English, 93% precision and 93% recall for French, and 92% precision and 89% recall for Greek.

11.2.3 Automatic Speech Recognition

In most state-of-the-art speech processing systems [6, 7], ASR is performed via a large vocabulary, speaker-independent, gender-independent, continuous speech recognizer. The recognizers typically use hidden markov models (HMM) with GMM for acoustic modelling (AM) and n-gram models estimated on large corpora for language modelling (LM).

The task for the ASR is to convert the information passed on to it by the SCD stage into time-tagged text. This is often done in a multi-stage approach by finding the most likely sequence of words, given the input audio stream. Each level of processing uses increasingly

complex acoustic and language models, thereby keeping a balance between model complexity, search space size and quality of the results. The AMs used by different stages are HMM-based mixture models with various degrees and kinds of tying (sharing of model parameters). They form a representation of the acoustic parameters (acoustics, phonetics, environment, audio source, speaker and channel characteristics, recording equipment etc.). The LM is an n-gram model with different types of back-off and smoothing. It contains a representation of the words included in a vocabulary, and of what words and sequences of words are likely to occur together or follow each other. The recognizer's vocabulary consists of several thousand wordforms and their pronunciation models. Using wordforms as opposed to words means that plural and singular forms (or any kind of declinational or conjugational forms) of a word are considered separate entities in the vocabulary.

ASR in CIMWOS uses a multi-stage approach to provide for flexibility and best use of resources. During *decoding*, in a first step, a *fast-match* is performed using the simplest (and thus fastest) models. Then a *detailed-match* follows on the results of fast-match. Finally a *rescoring pass* is applied to the results of the second stage. This staged approach allows for a dramatical reduction in search space while at the same time recognizer performance and accuracy are kept high. Evaluation of the CIMWOS ASR module on predefined test sets for English and french broadcast news (BNs) yielded word error rates (WERs) of 29% and 27%, respectively. For Greek BNs a 33% WER is reported. Figure 11.2 shows a sample portion of an XML file automatically created by the speech processing subsystem in CIMWOS.

```
<?xml version="1.0" encoding="UTF-8" ?>
<SpeechAnnotation project="CIMWOS">
<Header type="SpeechRecognitionMetadata">
  <Media id="RTBF_20011012_1930_News" xml:lang="fr">
  ….
  <Passage   id="p77 "      speaker="male    14"    gender ="male"
mediaRelIncrTimePoint="131864"         mediaIncrDuration="1420"
xml:lang="fr">
       <Word     id="w3541"      mediaRelIncrTimePoint="131864"
mediaIncrDuration="13" confidence="0.99">Dans</Word>
       <Word     id="w3542"      mediaRelIncrTimePoint="131878"
mediaIncrDuration="8" confidence="0.99">le</Word>
       <Word     id="w3543"      mediaRelIncrTimePoint="131887"
mediaIncrDuration="17" confidence="0.99">hall</Word>
       <Word     id="w3544"      mediaRelIncrTimePoint="131905"
mediaIncrDuration="10" confidence="0.99">de</Word>
       <Word     id="w3545"      mediaRelIncrTimePoint="131916"
mediaIncrDuration="48" confidence="0.99">départ</Word>
       <Word     id="w3546"      mediaRelIncrTimePoint="131965"
mediaIncrDuration="5" confidence="0.99">de</Word>
       <Word     id="w3547"      mediaRelIncrTimePoint="131971"
mediaIncrDuration="4" confidence="0.99">l'</Word>
       <Word     id="w3548"      mediaRelIncrTimePoint="131976"
mediaIncrDuration="108" confidence="0.99">aéroport</Word> …
   </Passage>
  ….
  </Media>
</SpeechAnnotation>
```

Figure 11.2 Segment of speech processing generated metadata

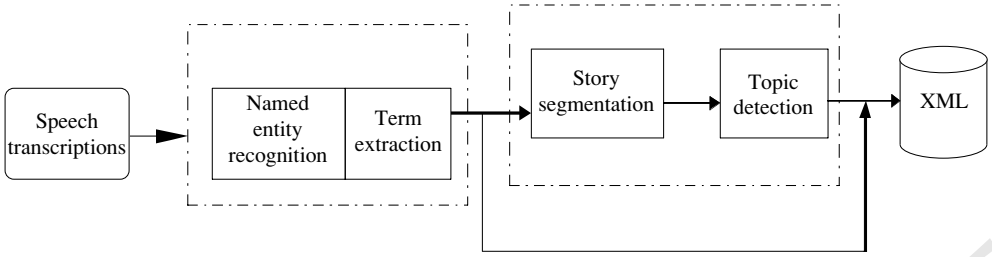

Figure 11.3 Text processing subsystem

11.3 Text Processing Subsystem

Following audio processing, text processing tools are applied on the textual data (speech transcriptions) produced by the speech processing subsystem, in an attempt to enable a retrieval system with the ability to answer questions like what topic a passage is about, or which organizations are mentioned. Named entity detection (NED), term extraction (TE), story segmentation (SD) and topic detection (TD) are tasks that can be included in a text processing pipeline like the one depicted in Figure 11.3.

11.3.1 Named Entity Detection and Term Extraction

The task of the NED module in the context of a multimedia retrieval system is to identify named locations, persons and organizations, as well as dates, percentages and monetary amounts, in the speech transcriptions automatically produced by the SPS. NED involves an initial preprocessing phase where sentence boundary identification, and part-of-speech tagging are performed on textual input. State-of-the-art NED implementations involve lookup modules that match lists of NEs and trigger words against the text, hand-crafted and automatically generated pattern grammars, maximum entropy modelling, HMM models, decision-tree techniques, SVM classifiers etc. [8].

As an example, the NED module for the Greek language in CIMWOS [9] is a combination of list-based matching and parsing with a grammar of rules compiled into finite-state transducers. In a collection of TV news broadcasts, this module identified 320 NEs, compared to the 555 identified by human annotators using a suitable annotation tool (Figure 11.4). There were 235 correct guesses, 85 false positives and 320 missed instances. Although promising precision scores were obtained for persons and locations, (Figure 11.5), lower recall in all NE types is due to:

- greater out-of-vocabulary (OOV) rate in the ASR for Greek;
- missing proper names in the vocabulary of the ASR engine;
- the diversity of domains found in broadcasts, leading to a large number of names and NED indicators that were not incorporated in the NED module's resources.

The TE task involves the identification of single or multi-word indicative keywords (index terms) in the output of the speech processing system. Systems for automatic term extraction using both linguistic and statistical modelling are reported in the literature [10]. Linguistic

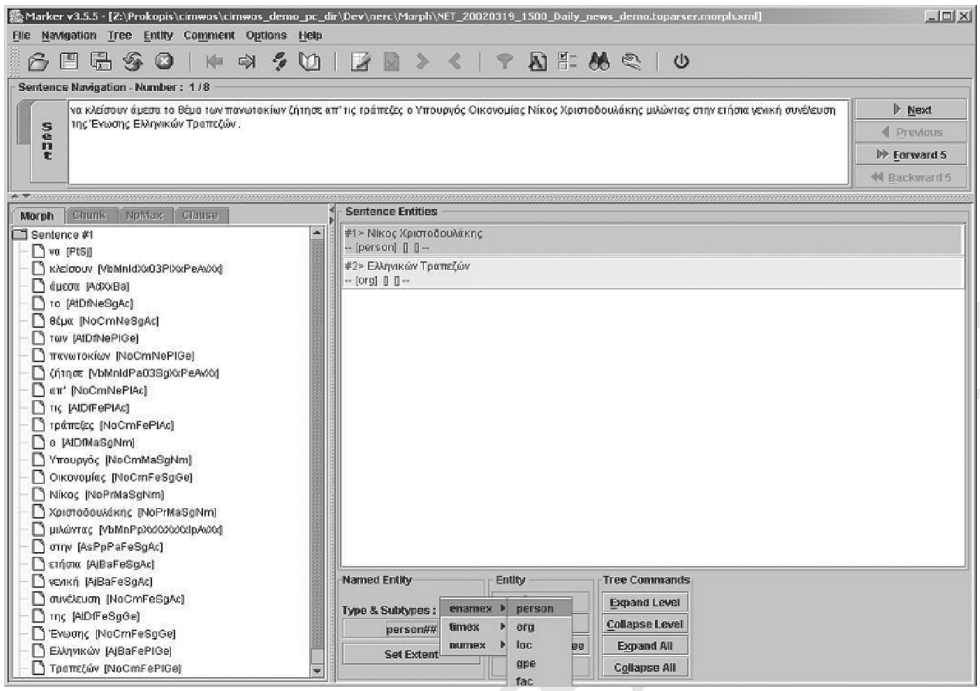

Figure 11.4 Named entity annotation tool

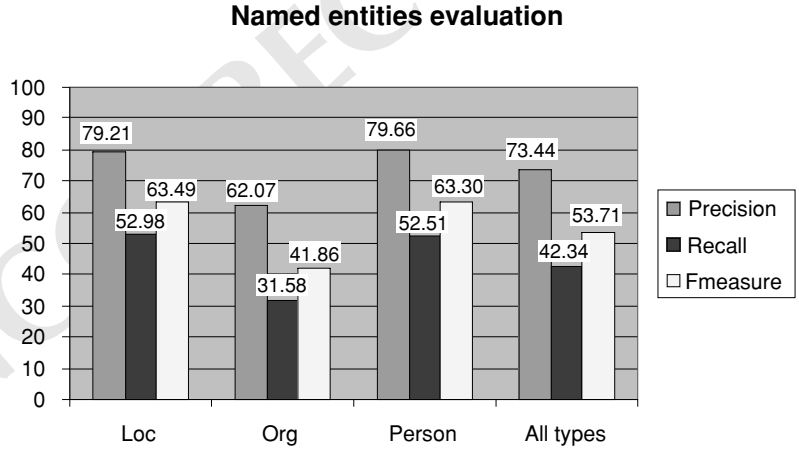

Figure 11.5 Performance of NED module for each NE type

processing is usually performed through an augmented term grammar, the results of which are statistically filtered using frequency-based scores.

The term extractor in CIMWOS was used to automatically identify terms in a testing corpus of manually annotated transcriptions, scoring a 60.28% recall and a 34.80% precision. To

obtain higher precision scores, stricter statistical filtering needs to be applied. Nevertheless, a relaxed statistical filtering like the one currently used produces higher recall measures, a factor which plays a more important role in an information retrieval context.

11.3.2 Story Detection and Topic Classification

Story detection and topic classification modules often employ the same set of models, trained on an annotated corpus of stories and their manually associated topics. The basis of these technologies is a generative, mixture-based hidden Markov model that includes one state per topic, as well as 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.

In the Informedia project at Carnegie Mellon University [11] each video document is partitioned into story units based on text, image and audio metadata jointly. Silence periods are identified and subsequently aligned with the nearest shot break indicating the boundary of a story unit. Text markers such as punctuation are taken into consideration in case closed caption is available. In the MAESTRO system [12] developed by SRI, topic boundaries are detected based on prosodic information extracted from the speech waveforms (pause and pitch patterns) combined with word usage statistics. The MITRE Broadcast News Navigator (BNN) incorporates a story segmentation module based on finite state machines [13].

11.4 Image Processing Subsystem

Image processing and understanding is an important field of research where close ties between academic and commercial communities have been established. In the multimedia content analysis framework, a typical image processing subsystem (IPS) like the one presented here consists of modules performing video segmentation and keyframe extraction, face detection and face identification, object identification and video text detection and recognition.

11.4.1 Video Segmentation and Keyframe Extraction

The segmentation of video sequences is a prerequisite for a variety of image processing applications. Video streams consist 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 aeroplane, or a fixed camera focusing on an anchorperson), while a shot cut is the point at which shots change within a video sequence. The video segmentation task involves detecting shot cuts and thus partitioning raw material into shots. Under the assumption that frames within a shot have a high degree of similarity for each shot, a few representative frames are selected, referred to as keyframes. Each keyframe represents 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 so that a single keyframe no longer suffices to describe the content of the entire shot. Keyframes contain most of the static information present in a shot, so that subsequent modules like face detection and object identification can focus on keyframes only without scanning the whole video sequence.

In order to detect shot cuts and select keyframes, different methods have been developed for measuring the differences between consecutive frames and applying adaptive thresholding on motion and texture cues. The performance of the CIMWOS video segmentation module was evaluated on some of the news broadcasts available in the project's collection. A recall rate of 99% and a precision of 97% were observed in the case of abrupt shot cuts. For smooth shot cuts, performance was lower and highly dependent on the video content, yet still comparable to the state of the art.

The performance of the keyframe selection is much harder to quantify, as there is no objective groundtruth available: what constitutes an appropriate keyframe is a matter of semantic interpretation by humans, and cannot be resolved on the basis of low-level image cues. Nevertheless, judging from comments by test users, the CIMWOS keyframe extraction module was indeed able to reduce the number of redundant keyframes while keeping important information available.

11.4.2 Face Detection and Face 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 face dimensions in the keyframe. Face detection is a challenging task, since several factors influence 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. It is one of the few attempts to recognize from a set of 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 experiments by neuroscientists and psychologists show that face recognition is a dedicated process in the human brain. This may have encouraged the view that artificial face identification systems should also be face-specific. Automatic face identification is a challenging task and has recently received significant attention. Rapidly expanding research in the field is based on recent developments in technologies such as neural networks, wavelet analysis and machine vision. Face identification has a large potential for commercialization in applications involving authentication, security system access, and advanced video surveillance. Nevertheless, in spite of expanding research, results are far from perfect, especially in uncontrolled environments, because of lighting conditions and variations, different 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 different facial expressions, lighting etc.) and extra-personal variation (between different individuals).

In CIMWOS, the face detection and face identification modules associate detected faces occurring in video streams with names [14] (Figure 11.6). Both modules are based on support

Figure 11.6 Face detection results

vector machine models trained on an extensive database of facial images with a large variation in pose and lighting conditions. Additionally, a semantic base consisting of important persons that should be identified has been constructed. 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 as 'unknown'.

In order to be able to compare the CIMWOS FD performance with other systems' results reported in the literature, the module was evaluated against the MIT/CMU dataset [15], where groundtruth information for frontal face detection is readily available. A correct detection rate of 90% (with 1 false detection in 10. 000 windows) was obtained. A qualitative evaluation on the CIMWOS news broadcasts database shows that similar results are obtained in that context, if the extremely small or out-of-plane rotated faces are ignored.

11.4.3 Video 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 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 usually of low resolution, of any colour or greyscale value (not always white), embedded in complex background. Experiments show that applying conventional OCR technology directly to video text leads to a 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.

Figure 11.7 Video text detection and recognition

In CIMWOS, the video text detection and recognition module [16, 17] is based on a statistical framework using state-of-the-art machine learning tools and image processing methods (Figure 11.7). It consists of four modules:

- Text detection, aiming at roughly and quickly finding image blocks that may contain a single line of text characters.
- Text verification, using a support vector machine model, to remove false alarms.
- Text segmentation, attempting to extract pixels from text images belonging to characters with the assumption that they have the same colour/greyscale value. This method uses a Markov random field model and an expectation maximization algorithm for optimization.
- 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 bigram language model.

The CIMWOS video text detector was evaluated on 40 minutes of video consisting of 247 text strings, 2899 characters and 548 words. Characters were correctly extracted and recognized in 94% of all cases, while a 92% recognition rate was reached as far as words are concerned. Precision showed that more than 95% of extracted characters correspond to the characters in the groundtruth. In the 2899 characters, 7% of characters were scene text characters and 93% were captions.

11.4.4 Object Identification

The object identification (OI) problem can be defined as the task of identifying a non-deformable man-made 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 3D objects and viewpoints is the 'appearance-based' one, which nevertheless requires a very large amount of example views and has fundamental problems in dealing with cluttered recognition views and occlusion [18]. A system capable of dealing with general 3D objects, from general viewing directions, is yet to be proposed.

Since 1998, a small number of systems 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 [19]. 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 the 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 spatio-temporal order inherent in the set of individual regions, and requires it to be present in the recognition view. In 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 recognized regions.

Thanks to the good degree of viewpoint invariance of the regions, and to the strong model and learning approach developed, the object recognition module in CIMWOS copes with 3D objects of general shape, requiring only a limited number of learning views [20]. Moreover, it can recognize objects from a wide range of previously unseen viewpoints, in possibly cluttered, partially occluded, views. Based on experiments conducted on the project's database (searching for 13 objects in 325 keyframes, on average, for each object), a 83.7% precision and a 96.4% recall were obtained.

11.5 Integration Architecture

Due to the prevalence of XML as a medium for data exchange between applications, all processing output in the three modalities (audio, image and text) can converge to a textual XML metadata annotation document following standard MPEG-7 descriptors. These annotations can be further processed, merged, synchronized and loaded into the multimedia database. A merging component amalgamates various XML annotations to create a self-contained object compliant with a database scheme.

The CIWMOS architecture follows an N-tier scenario by integrating a data services layer (storage and 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. Web services (SOAP) could be used as the integration protocol to access heterogeneous and loosely-coupled distributed technologies. The CIMWOS architecture is depicted in Figure 11.8.

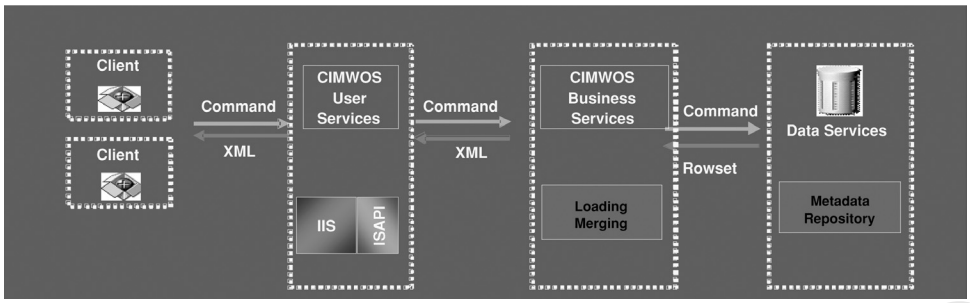

Figure 11.8 The CIMWOS architecture

11.5.1 Indexing, Search and Retrieval

The basic approach to indexing and retrieval is to apply speech, image and text technologies to automatically produce textual metadata and then to use information retrieval techniques on these metadata. The CIMWOS retrieval engine [21] is based on a weighted Boolean model. After query submission, the retrieval engine is invoked with the set of criteria combined flexibly with standard Boolean operators. A matching operation computes the similarity between the query and each passage to determine which ones contain the given set of combined query terms. The calculated similarity measure takes into consideration three different factors: metadata-level weights based on the overall precision of each multimedia processor, value-level statistical confidence measures produced by the engines and tf*idf scores for all textual elements. Finally, passages are ranked based on the result of similarity computation.

The basic retrieval unit is the passage (Figure 11.9), which has the role of a document in a traditional system. Passages are defined on the speech transcriptions—that is, on the

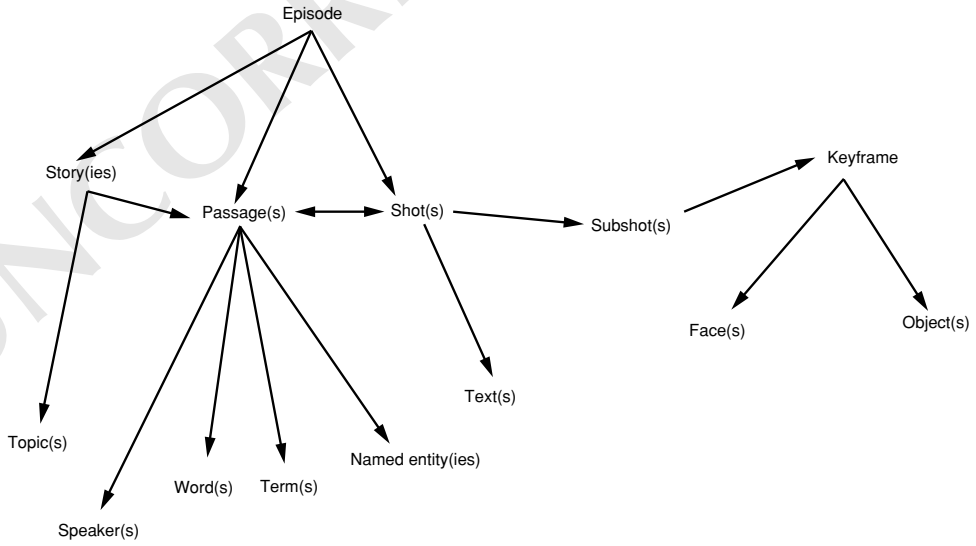

Figure 11.9 CIMWOS indexing schema

post-processed output of the ASR—therefore they correspond to some level of audio segmentation. Other processing streams, that is, text and video, will typically result in segments (stories and shots, respectively) not aligned with passages. This lack of correspondence with the original multimedia object has implications for both indexing and retrieval. The solution given must be based on the eventual use of the material, therefore the needs of the users should be taken into account. Referring to the example of a user searching for Putin co-occurring with Clinton on a particular crisis topic, it is necessary for a search and retrieval system to define the temporal extent over which the existence of both annotations counts as co-occurrence, meaning that the relevant time segment is to be retrieved. For news broadcasts, a news story is the appropriate conceptual unit. However, in the merged annotations one has access to processed outputs from the various streams, which do not map onto the desired unit in a well-defined manner. Thus the selection of passages is made to satisfy both practical constraints (availability, typical duration) and use requirements (to contain all relevant information and nothing irrelevant, as much as possible).

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 passage 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.

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 for someone to determine whether a clip meets his or her 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.

Multimedia information retrieval systems typically provide various visualized summaries of video content where the user can preview a specific segment before downloading it. In Informedia, information summaries can be displayed at varying detail, both visually and textually. Text summaries are displayed for each news story through topics and titles. Visual summaries are given through thumbnails, filmstrips and dynamic video skims [11]. In CIMWOS, while skimming a passage, the end user can view its associated metadata, the transcribed speech and results of all processing components, as well as a representative sequence of thumbnails. These thumbnails can act as indicative cues on how relevant the segment content is to the user query. Each thumbnail corresponds to a keyframe that is recognized and extracted by the video segmentation module. Results can also include more pieces of bibliographical information, such as the title of the video to which the passage belongs and the duration of the passage. Finally, the user can play the passage via streaming.

11.6 Evaluation

Apart from the evaluation of each module contributing metadata annotations, assessing the overall system response to user queries is of course of extreme importance for a multimedia information retrieval (MIR) system. Suggestions and guidelines on evaluation procedures have been put under testing in the framework of international contests. For example, the TREC 2002 Video track [22] included three sessions focusing on shot boundary detection, extraction of segments containing semantic features like ‘People’, ‘Indoors Location’ etc., and search for particular topics from a handcrafted list.

In what follows, a brief overview of the CIMWOS retrieval exercise is given. The CIMWOS system and its overall performance has been evaluated in two distinct phases. During the first phase, we tested video search and retrieval of passages relevant to a particular topic. We repeated the retrieval task during the second phase, this time on video material for which boundaries of relevant stories for each topic had been previously identified by human annotators. Our testing material consisted of Greek news broadcasts, produced by state and private TV networks between March 2002 and July 2003. For the first phase we used 15 videos (henceforth, Collection A) of a total duration of approximately 18 hours, captured in BETA SP and transcribed in MPEG-2 format. During the second phase we used 15 news broadcasts (henceforth, Collection B) that amounted to approximately 17 hours of video, captured via standard PC TV cards in MPEG-2 format.

The overall retrieval of the system was tested in both 'interactive' and 'manual' search modes. In interactive search, users are familiar with the test collection, and have full access to multiple interim search results. On the other hand, in the case of manual search, users with knowledge of the query interface but no direct or indirect knowledge of the search test set or search results are given the chance to translate each topic to what they believe to be the most effective query for the system being tested. In each phase, a user familiar with the test collection was responsible for generating a set of topics that s/he judged to be of interest, opting for topics that were represented in more than one news broadcast of the collection. After the list was finalized, the user had to manually locate all video sequences relevant to each topic, thus allowing developers to associate start and end timecodes with each topic.

During the search task, users were given the chance to translate each topic of the list to what they believed to be the most effective queries for the system being tested, using combinations of criteria like Terms, Named Entities and/or ASR Text. Users formed five queries, on average, for each topic. Using the HTML interface (Figure 11.10), they were able to browse the results in order to intuitively assess their overall satisfaction with the database results. Nevertheless, results were also saved in an XML format as in Figure 11.11, where the topic 'earthquake prediction' has been translated into a query comprising the word $\pi\rho\acute{o}\gamma\nu\omega\sigma\eta$ (*prediction*) and a name (the name of a Greek geology professor).

In their initial reactions when searching for a topic in the videos of Collection A, users reported that passages returned by the system were too short and fragmentary. The explanation was that passages were based on automatic segmentation by the ASR engine, based on hints from speaker changes. A different approach was taken while testing on Collection B, in order to produce more intuitive results for end users.

A user was again responsible for manually identifying relevant segments in each of the videos of the collection. These segments corresponded to the stories of each news broadcast. Following that, we aligned the ASR transcription with the manually identified stories. Each video segment was also assigned a description, derived from a list of topics created during the training phase of the Greek ASR module. Thus, while it was possible to have passages consisting of just one shot during the first phase, this counterintuitive phenomenon was absent in the semi-automatically annotated material of Collection B.

The XML files corresponding to each user's queries were assembled and tested against groundtruth data. In Table 11.1, we show results for both video collections. We also tested the system's response when we did not take into account results that scored less than 60% in the interface's ranking system. This filtering had a negative effect on the system's recall in the case of Collection B. Nevertheless, it significantly increased the system's precision with data from both sets.

Figure 11.10 CIMWOS HTML interface

11.7 Related Systems

Many multimedia indexing and retrieval systems have been developed over the past decade. Most of these systems aim at dynamically processing, indexing, organizing and archiving digital content [23, 24]. A brief overeview of the most relevant systems is given below.

News on demand (NOD) systems have made their appearance in the past decade, integrating image, speech and language technology, and taking advantage of cross-modal analysis in order to monitor news from TV, radio and text sources and provide personalized newscasts [13].

MIT's Photobook [25] was an early example that allowed retrieval from textures, shapes and human faces. Chabot [26], developed by UCBerkeley, is another early system. It provided a combination of text-based and colour-based access to a collection of photographs. The system was later renamed Cypress, and is now known as the DWR picture retrieval system [27], and has been incorporated into the Berkeley Digital Library project. Berkeley has continued similar research with its Blobworld software [28], which has one of the more sophisticated region segmentation functionalities. Noteworthy are recent developments at Berkeley, which allow finding of specific object classes like horses and naked people. VisualSEEK [29] was the first of a series of systems developed at Columbia University. A strong point was that queries were quite flexible in the spatial layout specifications that they could accept. WebSEEK [30] was a further development, aimed at facilitating queries over the web. Emphasis is on colour, but in another prototype called VideoQ [31], motion was added. Also, relative layouts of regions with specific colours or textures can be specified. This allows for some primitive

```
<?xml version="1.0" encoding="UTF-8"?>
 <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="ondemand delivery">
       <mediatitle>News      of      2002-05-24
          </mediatitle>
       <source duration="1:74:52,48" />
     </media>
     <thumbnails number="1" />
   </passage>
   …
  </results>
```

Figure 11.11 Query results in XML

forms of object detection, e.g. finding the American flag on the basis of the alternation of red and white stripes, with a blue region at the top left corner. The MARS project at the University of Illinois [32] put emphasis on control by the user, and introduced extensive relevance feedback mechanisms. Feature weights are set dynamically. An example of European

Table 11.1 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

CBIR technology is the Surfimage system from INRIA [33]. It has a similar philosophy to the MARS system, using multiple types of image features which can be combined in different ways, and offering sophisticated relevance feedback. The RSIA system co-developed by ETH Zurich and DLR [34] was dedicated to search in large satellite image databases. It supports remote access through the web and is best on texture features, learned as the user indicates positive and negative examples of the kind of regions searched for. The Netra system [35] uses similar cues to many of the foregoing systems, i.e. colour, texture, shape and spatial location information, but as in Blobworld a more sophisticated segmentation function is included. Synapse [36] is a rather different kind of system, in that it bases its similarity judgements on whole image matching. Query By Image Content [37] is a component of the IBM DB2 database product, performing image retrieval based on example images and selected colour and texture patterns.

Convera [38] provides scene change detection and can send audio tracks to any SAPI-compliant recognition engine. It also provides a closed captioning/teletext extractor and an SMPTE timecode [39] reading module, but these are more 'decoders' than true 'recognition engines', as this information is already available in digital form. On its part, Virage [40] claims to be more comprehensive, including face recognition, on-screen text recognition, speech and speaker, as well as intelligent segmentation.

11.8 Conclusion

Multimedia information retrieval systems address real needs of multimedia producers, archivists and others who need to monitor and/or index audiovisual content. By utilizing vast amounts of information accumulated in audio and video, and by employing state-of-the-art speech, image and text processing technologies, an MIR system can become an invaluable assistant in efficient reuse of multimedia resources, and in reducing new production expenses.

These systems open new possibilities and provide enabling technology for novel application areas and services. Use of markup languages (MPEG-7, MPEG-21) for annotation of audiovisual content can promote standardization and unification of access procedures to large archives and multiple repositories, thus assisting the rise of multimedia digital libraries, and improving the working conditions and productivity of people involved in media and television, video, news broadcasting, show business, advertisement and any organization that produces, markets and/or broadcasts video and audio programmes.

Acknowledgements

Project CIMWOS was supported by shared-cost research and technological development contract IST-1999-12203 with the European Commission. Project work was done at the Institute for Language & Speech Processing, Athens, Greece (coordination, integration, text processing, search and retrieval), Katholieke Universiteit Leuven, Belgium (video segmentation, face detection), Eidgenoessische Technische Hochschule Zurich, Switzerland (scene and object recognition), Sail Labs Technology AG, Vienna, Austria (audio processing, speech recognition, text processing), Canal+ Belgique, Brussels, Belgium (user requirements and validation) and Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny, Switzerland (speech recognition, video text detection and recognition, face recognition).

References

[1] H. Papageorgiou, A. Protopapas, CIMWOS: a multimedia, multimodal and multilingual indexing and retrieval system. In E. Izquierdo (ed.) *Digital Media Processing for Multimedia Interactive Services, Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services*, Queen Mary, University of London, 9–11 April 2003, pp. 563–568. World Scientific, Singapore, 2003.

[2] Y. Wang, Z. Liu, J.-C. Huang, Multimedia content analysis: using both audio and visual clues. *IEEE Signal Processing Magazine*, **17**(6), 12–36, 2000.

[3] Z. Liu, Y. Wang, T. Chen, Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, **20**(1), 61–79, 1998.

[4] J.-L. Gauvain, R. De Mori, L. Lamel, Advances in large vocabulary speech recognition. *Computer Speech and Language*, **16**(1), 1–3, 2002.

[5] J.-L. Gauvain, L. Lamel, G. Adda, Audio partitioning and transcription for broadcast data indexation. *Multimedia Tools and Applications*, **14**, 187–200, 2001.

[6] F. Kubala, J. Davenport, H. Jin, D. Liu, T. Leek, S. Matsoukas, D. Miller, L. Nguyen, F. Richardson, R. Schwartz, J. Makhoul, The 1997 BBN BYBLOS System applied to Broadcast News Transcription. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 8–11 February 1998. National Institute of Standards and Technology, Gaithersburg, MD, 1998. Available at: http://www.nist.gov/speech/publications/darpa98/pdf/eng110.pdf.

[7] J.L. Gauvain, L. Lamel, G. Adda, Transcribing broadcast news for audio and video indexing. *Communications of the ACM*, **43**(2), 64–70, 2000.

[8] E.F. Tjong Kim Sang, Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In D. Roth, A. van den Bosch (eds) *CoNLL-2002, Sixth Conference on Natural Language Learning*, Taipei, Taiwan, 31 August –1 September 2002. Available at: http://cnts.uia.ac.be/conll2002/proceedings.html.

[9] I. Demiros, S. Boutsis, V. Giouli, M. Liakata, H. Papageorgiou, S. Piperidis, Named entity recognition in Greek texts. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000*, Athens, Greece, 31 May–2 June 2000, pp. 1223–1228. ELRA, 2000.

[10] C. Jacquemin, E. Tzoukermann, NLP for term variant extraction: synergy between morphology, lexicon and syntax. In T. Strzalkowski (ed.) *Natural Language Information Retrieval*. Kluwer, Dordrecht, 1999.

[11] A. Hauptmann, M. Smith, Text, speech, and vision for video segmentation: the Informedia Project. In *AAAI Fall 1995 Symposium on Computational Models for Integrating Language and Vision*. American Association for Artificial Intelligence, 1995.

[12] SRI Maestro Team, MAESTRO: conductor of multimedia analysis technologies. *Communications of the ACM*, **43**(2), 57–63, 2000.

[13] A. Merlino, D. Morey, M. Maybury, Broadcast news navigation using story segments. In *Proceedings of the Fifth ACM International Conference on Multimedia*, Seattle, WA, 9–13 November 1997, pp. 381–391. ACM, Seattle, WA, 1997.

[14] F. Cardinaux, C. Sanderson, S. Marcel, Comparison of MLP and GMM classifiers for face verification on XM2VTS. In J. Kittler, M.S. Nixon (eds) *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, 9–11 June 2003. Springer-Verlag, Berlin, 2003.

[15] The Combined MIT/CMU Test Set with Ground Truth for Frontal Face Detection, http://vasc.ri.cmu.edu/idb/html/face/frontal_images/index.html.

[16] D. Chen, H. Bourlard, J.-Ph. Thiran, Text identification in complex background using SVM. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001, vol. 2, pp. 621–626. IEEE, 2001.

[17] J.M. Odobez, D. Chen, Robust video text segmentation and recognition with multiple hypotheses. In *Proceedings of the International Conference on Image Processing, ICIP 2002*. IEEE, 2002.

[18] V. Ferrari, T. Tuytelaars, L. Van Gool, Wide-baseline muliple-view correspondences. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition ICIP 2003*, Madison, WI, June 2003. IEEE, 2003.

[19] T. Tuytelaars, A. Zaatri, L. Van Gool, H. Van Brussel, Automatic object recognition as part of an integrated supervisory control system. In *Proceedings of IEEE Conference on Robotics and Automation, ICRA00*, San Francisco, CA, April 2000, pp. 3707–3712. IEEE, 2000.

[20] V. Ferrari, T. Tuytelaars, L. Van Gool, Real-time affine region tracking and coplanar grouping. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001, vol. 2. IEEE, 2001.

[21] H. Papageorgiou, A. Protopapas, T. Netousek, Retrieving video segments based on combined text, speech and image processing. In *Proceedings of the Broadcast Engineering Conference*, April 2003, pp. 177–182. National Association of Broadcasters, 2003.
[22] A.F. Smeaton, Paul Over, The TREC-2002 Video track report. In E.M. Voorhees, L.P. Buckland (eds) *The Eleventh TExt Retrieval Conference (TREC 2002)*, Gaithersburg, MD, 19–22 November 2002. NIST Special Publication 500-251. National Institute of Standards and Technology, Gaithersburg, MD, 2002. Available at: http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
[23] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, A. Zakhor, Applications of video-content analysis and retrieval. *IEEE Multimedia*, **9**(3), 42–55, 2002.
[24] *IEEE Multimedia*, Special issue: Content-Based Multimedia Indexing and Retrieval, **9**(2), 18–60, 2002.
[25] A. Pentland, R.W. Picard, S. Sclaroff, Photobook: content-based manipulation of image databases. *International Journal of Computer Vision*, **18**(3), 233–254, 1996.
[26] V.E. Ogle, M. Stonebraker, Chabot: retrieval from a relational database of images. *IEEE Computer*, **28**(9), 164–190, 1995.
[27] DWR photo database, http://elib.cs.berkeley.edu/photos/dwr/about.html.
[28] Blobworld, http://elib.cs.berkeley.edu/photos/blobworld.
[29] J.R. Smith, S.-F. Chang, Tools and techniques for color image retrieval. In I.K. Sethi, R.C. Jain (eds) *Storage and Retrieval for Still Image and Video Databases IV*, SPIE vol. 2670, pp. 426–437. SPIE, 1996.
[30] J.R. Smith and S.-F. Chang, Image and video search engine for the world wide web. In I.K. Sethi, R.C. Jain (eds) *Storage and Retrieval for Image and Video Databases V*, SPIE vol. 3022, pp. 84–95. SPIE, 1997.
[31] S.F. Chang, W. Chen, H.J. Meng, H. Sundaram, D. Zhong, VideoQ: an automated content-based video search system using visual cues. In *Proceedings of International Multimedia Conference*, Seattle, WA, November 1997, pp. 313–324. Addison-Wesley, Reading, MA, 1997.
[32] T.S. Huang, S. Mehrotra, K. Ramchandran, Multimedia Analysis and Retrieval System (MARS) project. In P.B. Heidorn, B. Sandore (eds) *Proceedings of the 33rd Annual Clinic on Library Application of Data Processing: Digital Image Access and Retrieval*, Urbana, IL, March 1996, pp. 100–117. University of Illinois, 1997.
[33] C. Nastar, M. Mitschke, C. Meilhac, N. Boujemaa, Surfimage: a flexible content-based image retrieval system. In *Proceedings of International Multimedia Conference*, Bristol, 12–16 September 1998, pp. 339–344. Addison-Wesley, Harlow, 1998.
[34] K. Seidel, M. Datcu, G. Schwarz, L. Van Gool, Advanced remote sensing information system at ETH Zurich and DLR/DFD Oberpfaffenhofen. In IEEE International Geoscience and Remote Sensing Symposium IGARSS '99, Hamburg, Germany, pp. 2363–2365. IEEE, 1999.
[35] W.Y. Ma, B.S. Manjunath, NeTra: a toolbox for navigating large image databases. In *Proceedings of International Conference on Image Processing*, Santa Barbara, CA, October 1997, vol. 1, pp 568–571. IEEE, 1997.
[36] R. Manmatha, S. Ravela, Syntactic characterization of appearance and its application to image retrieval. In B.E. Rogowitz, T.N. Pappas (eds) *Human Vision and Electronic Imaging II*, SPIE vol. 3016, pp. 484–495. SPIE, 1997.
[37] Query By Image Content, http://wwwqbic.almaden.ibm.com.
[38] Convera, http://www.convera.com.
[39] Society of Motion Picture and Television Engineers, http://www.smpte.org.
[40] Virage, http://www.virage.com.