Audio-visual content analysis in P2P: the SAPIR approach

Walter Allasia
EURIX Group
Torino, Italy
allasia@eurixgroup.com

Fabrizio Falchi
ISTI-CNR
Pisa, Italy
fabrizio.falchi@isti.cnr.it

Francesco Gallo
EURIX Group
Torino, Italy
gallo@eurixgroup.com

Mouna Kacimi
Max-Planck-Institut für Informatik
Saarbrücken, Germany
mkacimi@mpi-inf.mpg.de

Aaron Kaplan
Xerox Research Centre Europe
Meylan, France
aaron.kaplan@xrce.xerox.com

Jonathan Mamou
IBM Research Lab
Haifa, Israel
mamou@il.ibm.com

Yosi Mass
IBM Research Lab
Haifa, Israel
yosimass@il.ibm.com

Nicola Orio
University of Padova
Padova, Italy
orio@dei.unipd.it

Abstract

Content based search in audio-visual collections requires media specific analysis for extracting low level features to be efficiently indexed and searched. We present the SAPIR media framework for analyzing digital content and representing the extracted features in a common schema. The framework contains splitters of compound objects to simple objects to deal with complex media like videos, using image and speech analyzers. The extracted features are then merged into a common representation. We report usage of this framework in the SAPIR demo.

1. Introduction

Web search for audio-visual content such as images, music, animations, and videos is limited today to associated text and metadata annotations. As a result the effectiveness of a retrieval session is highly dependent on the manual tagging of audio-visual content done by non professional users. On the other hand, real content-based audio-visual search requires media specific understanding for extracting low level features such as color histograms and textures for images, phonemes for speech, or video segmentation and low level features extracted from the segments. The use of low level features only, without the semantic annotations, generally results into bad search performance; the combination of low level features and semantic annotations leads to the best results.

The European project SAPIR (Search in Audio-visual content using Peer-to-peer Information Retrieval) aims at developing a large-scale, distributed Peer-to-Peer infrastructure that will make it possible to search in audio-visual content by a 'Query by example' paradigm, where the user can supply the query in the form of images, speech, etc., enhanced by metadata annotations. SAPIR aims at a complete end-to-end solution for large scale search in audio-visual content that includes media analysis framework, scalable and distributed P2P index structures supporting similarity search and support for multiple devices embedding social networking in a trusted environment.

The SAPIR vision is illustrated in Figure 1. In this paper we focus on the SAPIR media analysis framework for analyzing compound objects and representing the extracted features in a common schema based on MPEG-7 [10]. Examples of compound objects can be web pages containing images and textual metadata (e.g., Flickr[1]) or containing videos and metadata (e.g., YouTube[2]). Compound objects are split into simple objects and for each simple media type (speech, text, image) we implemented specific analyzers (*annotators*), combined for dealing with complex media like videos, where images and speech, and possibly music, are separately analyzed. Finally the features extracted by each annotator are merged into a common schema. The implemented annotators are able to exploit the P2P architecture, where each component can run on a different peer.

[1] http://www.flickr.com

[2] http://www.youtube.com

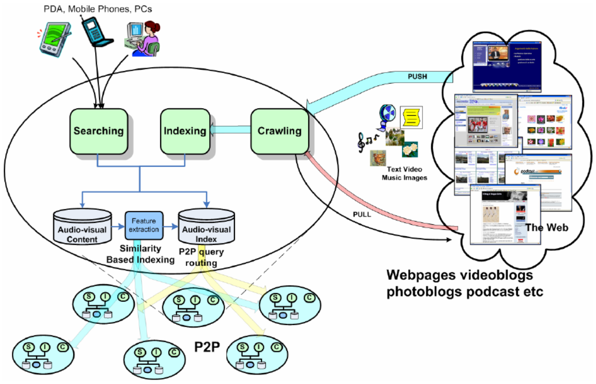

Figure 1. SAPIR components and functions

2. SAPIR Architecture

SAPIR implements a flexible architecture in order to deal with the representation, organization and storage of several different data types. We define four main architectural components reflecting a common peer user case, as shown in Figure 2: the peer joins the P2P system and its communication infrastructure by means of the *Networking* component; the *Content Management* component is responsible for the content analysis (feature extraction) of the digital items that the peer wants to share (text documents, images, audio or video recordings); the *P2P Indexing* component is responsible for publishing (indexing) the outcomes of the content analysis; search requests for similar or other contents are performed by the *P2P Search* component. The interested reader can find more details about indexing, search and networking components in the SAPIR web site[3].

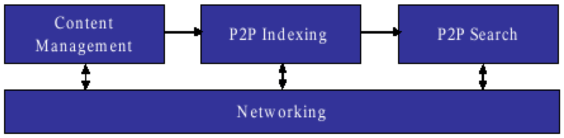

Figure 2. SAPIR Software Components

The Content Management component includes content injection and metadata enrichment, providing capabilities for the analysis of different content types (text, image, audio and video). Feature extraction can be based on text descriptions or media specifications including contours and color distributions in images, video segmentation into scenes and selection of characteristic frames, speech-to-text extraction, melody and rhythm extraction. The SAPIR approach aims at making use of standards for metadata representation as well as content analysis methods: metadata are expressed with an XML format derived from MPEG-7, while the UIMA framework [1] has been chosen to analyze content and extract its relevant features. Figure 3 depicts the basic principle of the SAPIR feature extraction framework.

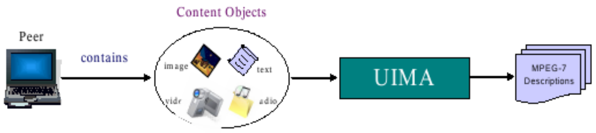

Figure 3. Feature Extraction Flow

Digital contents in SAPIR are defined by *Content Objects*, that can represent text or markup documents or multimedia objects such as images, audio, and video clips. Content Objects are identified by a unique identifier (URI) and implement a *Composite Design Pattern* where Compound Content Objects can be made of Simple Content Objects that represent the leaves of our schema. The Content Management component is responsible for analyzing Content Objects and extracting the features which are properties of Content Objects. Example features are ColorHistogram or Contours for images, rhythm or pitch sequence for music and terms (single words, possibly lemmatized) for textual Content Objects. But even for text, more advanced features such as grammatically tagged words, phrases, or named entities are conceivable, and web pages like Flickr have annotations (social tags) and metadata (e.g., GPS coordinates) as explicit features. SAPIR does not aim to provide a comprehensive catalog but proposes a framework for extensible and configurable feature sets.

3. Content Analysis

In this Section, we describe the media content analysis to extract the features used for indexing and search. Each Content Object may contain a combination of text, speech, music, images and video. For each content, the system generates an XML, which describes the content structure and the extracted features for each of its parts. The XML format derives from MPEG-7, with some extensions that will be proposed as a contribution by SAPIR to the evolution of the standard. The software has been implemented using the Unstructured Information Management Architecture (UIMA) platform [1] described below. Further details about the processing of each data type and the XML representation of the results can be found on the SAPIR web site. Since the content analysis is very challenging and can be quite resource consuming, it is possible to exploit the P2P network to have content analysis done on one peer while indexing and search

[3] http://www.sapir.eu

done on other peers. Moreover, it is even possible to split the content analysis of compound objects to several peers where each peer performs only a specific step of the content analysis (e.g. speech-to-text extraction or video decomposition). In the following we describe the components used for the analysis of each data type.

3.1. UIMA: Unstructured Information Management Architecture

Unstructured Information Management applications are software systems that analyze large volumes of unstructured information to discover knowledge that is relevant to an end user. UIMA [1] is an architecture and software framework for analysing unstructured content such as text, audio and video, for creating, discovering, composing and deploying a broad range of multi-modal analysis capabilities and integrating them with search technologies. The architecture is undergoing a standardization effort, referred to as the UIMA specification within OASIS[4]. Each component implements interfaces defined by the framework and provides self-describing metadata via XML descriptor files. The framework manages these components and the data flow between them. UIMA additionally provides capabilities to wrap components as network services, and can scale to very large volumes by replicating processing pipelines over a cluster of networked nodes. The SAPIR content analysis system is made up of several components, each specialized in processing a particular type of Content Object and at the lowest level annotators process simple Content Objects.

3.2. Speech Recognition

We use an Automatic Speech Recognition (ASR) system for transcribing speech data. For best recognition results, an acoustic model and a language model are trained in advance on data with similar characteristics. The ASR system generates word lattices. We use the compact representation of word lattice called *word confusion network* (WCN) proposed by [5, 3]. Each edge is labeled with a word hypothesis and its *posterior probability*, *i.e.*, the probability of the word given the signal. One of the main advantages of WCN is that it also provides an alignment for all of the words in the lattice. As stated by [5], although WCN are more compact than word lattices, in general the 1-best path obtained from WCN has a better word accuracy than the 1-best path obtained from the corresponding word lattice. Word decoding is also converted to its phonetic representation using the pronunciation dictionary of the ASR system. In addition, we generate phonetic output using a word-fragment decoder, where word-fragments are defined as variable-length sequences of phones [13]. The decoder generates 1-best word-fragments, that are then converted into the corresponding phonetic strings. We believe that the use of word-fragments can lead to a better phone accuracy compared to a phone-based decoder because of the strong constraints that it implies on the phonetic string.

3.3. Text processing

At a minimum, a textual Content Object is tokenized into words, and each word is mapped to a canonical form, which involves mapping capital letters to lowercase as well as stemming or lemmatization. For some applications, higher-level processing such as information extraction [12, 2] or summarization may be performed. For example, expressions in the text that denote people, places and dates can be annotated, and the document can be enriched with an automatically-generated summary.

3.4. Image analysis

We extract several MPEG-7 *visual descriptors* from each image. A visual descriptor characterizes a particular visual aspect of the image. They can be, therefore, used to identify images which have a similar appearance. Visual descriptors are represented as vectors, and the MPEG group proposed a distance measure for each descriptor to evaluate the similarity of two objects [7]. We have chosen five MPEG-7 visual descriptors [7, 4]: *Scalable Color*, *Color Structure*, *Color Layout*, *Edge Histogram* and *Homogeneous Texture*.

3.5. Music analysis

The first step in music processing is the automatic extraction of high level features, which are conveyed by music Content Objects either in scores or audio recording formats. The main content descriptors are rhythm and melody of the leading voice, which are automatically transcribed depending on the format: classification techniques are applied to symbolic scores to select the main melodic channel, while pitch tracking techniques are applied to digital recordings to highlight the main melody. All these techniques have been tested on a large dataset of music files. The melodic and rhythmic information is quantized, in order to take into account local variations of pitch and tempo. The second step is related to the segmentation of the extracted melody in musical lexical units, which are used as content descriptors. To this end, pattern analysis is applied to the sequence of notes forming the melody, highlighting all different patterns of a length from 3 to 6 notes. These thresholds have been experimentally evaluated [11]. Patterns are computed taking into account different melodic features, namely rhythm, pitch contour, and the combination of the two. The final step regards a suitable coding of the patterns, which differs

[4] http://www.oasis-open.org

according to the features to be represented. Given that all features undergo quantization, a textual representation will be used to describe the patterns, assigning a musical meaning to the symbols from a known alphabet [8]. The size of the alphabet depends on the choice of the quantization step. Commonly used values are 15 symbols for melodic intervals and 9 symbols for rhythm.

The relative frequency within a music content is computed for each pattern and each feature, as well as the time location of all its occurrences. The latter information can be used for fast access to relevant excerpts of the retrieved items. Music Content Objects are thus described by a set of patterns, their relative frequency, and a list of timestamps. Patterns are represented by sequences of symbols, from a predefined set (i.e., letters of English alphabet), the relative frequency is a floating point number, while timestamps are expressed in hundreds of seconds. The collection is indexed using a weighting scheme that takes into account the relative frequency of the patterns. This approach allows us to perform retrieval using exact match techniques between patterns, with positive effects on efficiency.

3.6. Video processing

The video processing performed in SAPIR follows the specifications of UIMA[5]. It currently works in three steps: the input video is analyzed and segmented into shots; the representative key frames for each shot are extracted and analyzed; the clustering algorithm creates clusters of shots based on key frames. Steps 1 and 2 are compliant to the guidelines on video analysis provided by the Prestospace [9] project. The multi-level temporal decomposition of the input video is represented by using the standard MPEG-7 schema. The video is decomposed at the first level using shots and clusters and at the second level using keyframes. Each element (keyframe, shot or cluster) is unambiguosly identified inside the MPEG-7 representation and all the keyframes are stored on the file system as JPG files. The shot segments are temporal units which can be obtained using different criteria, such as a fixed temporal interval or variable intervals extracted by video scene detection. The keyframes are the representative images for a particular shot (more than one keyframe per shot can be extracted) and are represented using the standard MPEG-7 *StillRegion* type. Finally all the extracted keyframes for all shots are compared and grouped according to their features. Each group (named cluster) contains a set of shots (which the grouped keyframes belong to) and a reference image for the whole cluster chosen among the similar keyframes for that group. Only the keyframes chosen to represent each cluster will be used for indexing and searching.

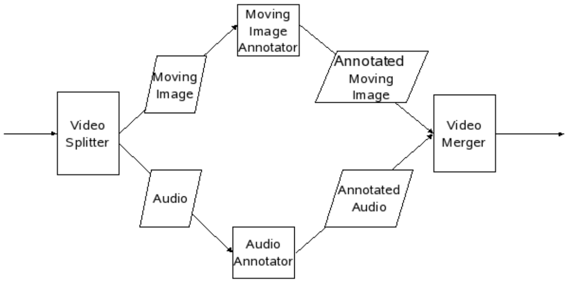

Figure 4. Video Annotator schema

The video analysis module is implemented as an UIMA *Annotator*, as shown in Figure 4. The video is split into moving image and audio parts, which are analyzed separately. The moving image part is further split into frames, which are processed by the same image annotator used for still images (not shown) in order to extract visual descriptors for each representative image. Once the annotators have completed the processes, the annotations are merged in order to be combined in a global metadata representation of the compound multimedia object. Each step of the video analysis described above can be run on a different peer.

4. Experiments

In the following we briefly describe the testbed collections and the experimental results related to the different digital media types in SAPIR. While the full demo is not yet publicly available, several demos related to the technologies used in SAPIR can be found on the project site.

Photo sharing sites provide a significant amount of additional metadata about hosted photos. The photo file contains information about the camera used, the time when it was taken, the aperture, the shutter speed, etc. Moreover, each photo comes with the name of the author, its title, a description, often with user-provided tags. Sometimes also richer information is available such as comments of other users on the photo, the GPS coordinates of the location where the photo was taken, the number of times it was viewed, etc. We decided to crawl Flickr, one of the most popular photo sharing sites providing permanent and centralized access to user-provided photos. This approach has several advantages: image quality, collection stability, legal issues and rich metadata. Flickr provides the richest additional metadata and an efficient API to access its content.

Within the SAPIR project a collection named CoPhIR[6] (COntent-based Photo Image Retrieval) has been developed to make significant tests on CBIR scalability. CoPhIR is go-

[5] http://incubator.apache.org/uima

[6] http://cophir.isti.cnr.it

ing to be made available to the research community to try and compare different indexing technologies for similarity search, with scalability being the key issue. When completed, the CoPhIR test-bed will contain standard MPEG-7 visual descriptors extracted from 100 millions high-quality public photos crawled from the Flickr photo-sharing site.

In SAPIR both visual descriptors and text are indexed. Thus both content-based similarity and text searches can be performed. An IR-style query language for multimedia content based retrieval has been developed for SAPIR. It exploits the XML representation of MPEG-7, extending the *XML Fragments* query language that was originally designed as a Query-By-Example for text-only XML collections [6]. In SAPIR it is also possible to perform complex similarity search combining results lists obtained using distinct features, GPS information and text. To this aim, state of the art algorithms for combining results are used.

Up to now we have conducted experiments over ten milion photos, performing similarity, text and combined searches. Experiments revealed the scalability of our approaches with response times from 0.2 second (for similarity search) to 2 seconds (for combined searches).

The experiments on music retrieval have been carried out on different collections, in order to test the applicability of the approach to different formats and different music genres. An initial collection of pop songs in Midi format of about 2000 items has been used to tune the parameters of the feature analysis, melodic segmentation and quantization. Additional experiments have been carried out on a collection of orchestral music, which contains about 18000 audio recordings. In order to carry out extensive testing on efficiency, thanks to a collaboration with a big content provider, we are currently collecting a large dataset of popular music of about 250000 audio files.

5. Conclusions

In this paper we presented the content analysis framework developed within the SAPIR project, aiming at developing a large-scale distributed P2P infrastructure for content-based search. The analysis framework and the approach used for each media type have been described. Finally we reported about several available demos based on different technologies and large samples, which will be integrated in the full public SAPIR demo.

6. Acknowledgments

This work was partially supported by the SAPIR project, funded by the European Commission under IST FP6 (Sixth Framework Programme, Contract no. 45128).

References

[1] D. Ferrucci and A. Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[2] C. Hagège, Ágnes Sándor, and A. Schiller. Linguistic processing of biomedical texts. In *Proceedings of PorTAL 2002, Portugal for Natural Language processing*, Taro, Portugal, 2002.

[3] D. Hakkani-Tur and G. Ricciardi. A General Algorithm for Word Graph Matrix Decomposition. In *Proceedings of the IEEE Internation Conference on Acoustics, Speech and Signal Processing (ICASSP), Hong-Kong*, pages 596–599, 2003.

[4] ISO/IEC. Information technology - Multimedia content description interfaces. Part 6: Reference Software, 2003. 15938-6:2003.

[5] E. B. Lidia Mangu and A. Stolcke. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400, 2000.

[6] J. Mamou, Y. Mass, M. Shmueli-Sheuer, and B. Sznajder. Query language for multimedia content. In *Procedding of the Multimedia Information Retrieval workshop held in conjunction with the 30 th Annual International ACM SIGIR Conference 27 July 2007, Amsterdam*, 2007.

[7] B. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[8] M. Melucci and N. Orio. Combining melody processing and information retrieval techniques: Methodology, evaluation, and system implementation. *Journal of the American Society for Information Science and Technology*, 55(12):1058–1066, 2004.

[9] A. Messina, L. Boch, G. Dimino, W. Bailer, P. Schallauer, W. Allasia, M. Groppo, M. Vigilante, and R. Basili. Creating Rich Metadata in the TV Broadcast Archives Environment: the Prestospace Project. In *AXMEDIS'06: Proceedings of the 2nd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*, pages 193–200, 2006.

[10] MPEG-7. Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002, 2002.

[11] G. Neve and N. Orio. A comparison of melodic segmentation techniques for music information retrieval. In *Proceedings of the European Conference on Digital Libraries*, pages 49–56, 2005.

[12] A. Rebotier, Ágnes Sándor, S. Voyatzi, T. Nakamura, C. Martineau, T. Delevallade, P. Capet, and J. Jacquelinet. Intelligent awareness: event extraction, information evaluation and risk assessment. In *3rd Language and Technology Conference*, Poznan, Poland, October 2007.

[13] O. Siohan and M. Bacchiani. Fast Vocabulary Independent Audio Search using Path based Graph Indexing. In *Proceedings of Interspeech, Lisbon*, pages 53–56, 2005.