System Architecture for Indexing Regions in Keyframes

Xavier Giró
Technical University of Catalonia (UPC)
Jordi Girona 1-3
Barcelona, Catalonia (Spain)
xavier.giro@upc.edu

Ferran Marqués
Technical University of Catalonia (UPC)
Jordi Girona 1-3
Barcelona, Catalonia (Spain)
ferran@gps.tsc.upc.edu

ABSTRACT
This paper describes the design of an indexing system for a video database. The system uses region-based manual annotations of keyframes to create models to automatically annotate new keyframes also at the region level. The presented architecture includes user interfaces for training and querying the system, internal databases to manage ingested content and modelled semantic classes, as well as communication interfaces to allow the system interconnection. The scheme is designed to work as a plug-in to an external Multimedia Asset Management (MAM) system.

Categories and Subject Descriptors
H.2.4 [**Database Management**]: Systems

General Terms
Design

Keywords
indexing, architecture, MAM, keyframe, region

1. INTRODUCTION
Content producers and TV broadcasters have migrated during the last decade from analogue to digital systems, turning their shelves full of tapes into large multimedia databases called Media Assets Management (MAM) systems. These systems have enormously improved their capacity of ingesting, producing and delivering new content but, at the same time, have increased their need of generating metadata associated to it. This situation, combined with the high cost of manual annotation, has raised the interest in automatic content indexing.

This paper presents the architecture of an indexing system aimed at helping documentalists in their task of indexing the growing amount of audiovisual content managed by MAMs. The system is designed as a plug-in for a MAM and acts as an external annotator that generates keywords for the visual content. The analysis is performed on keyframes previously extracted from each video asset and is based on image processing techniques that process keyframes at the region level. The design includes the elements necessary to train the system in generating new keywords automatically.

The document is divided in two parts. Firstly, the description of the six basic elements in the architecture and, secondly, the description of three basic user operations that can be performed with these elements. In all cases, the extraction of keyframes from video sequences is considered to be performed at a previous stage.

2. ARCHITECTURE
The system is composed of six different elements which may run in different and remote machines.

2.1 Content Database
All keyframes from video assets stored in the MAM are analyzed by the indexing system to create their region-based representation. This pre-processing corresponds to the extraction of visual descriptors from the data, a task that typically implies an important computational effort. As these features will be repeatedly assessed by different queries in the future, they are generated at ingest time and stored for future use in the Content DB

2.2 Knowledge Database
The keywords used by the MAM to index its contents are internally treated by the system as descriptions of another entity called semantic class, which can represent either an object, person, location or event. All considered classes are stored in a Knowledge DB and organized in one or several ontologies that represent their relationships. In addition to the mentioned keywords, this database also mantains an index of annotated instances for each semantic class .

2.3 Model Database
Each semantic class has one or more data models built after a training process. These models contain the classifier parametes and are typically learnt from the manual annotations of a small portion of the content database.

2.4 Instance Database
Each annotated instance is stored in the Instances DB, specifying the video asset, keyframe, semantic class and degree

of confidence on the annotation. Annotations can be generated manually or automatically and have an author associated, whether a human (manual annotations) or a machine (automatic annotations).

2.5 User Interfaces

Users interact with the system through graphical interfaces from which they can perform the following basic operations: annotation of key-frames at region-level, revision of automatic annotations generated by the system, management of semantic ontologies and search based on text or visual examples.

2.6 Communication Interfaces

Each element in the system includes a communication interface that establishes and controls the data transmissions. Two different types of protocols are used depending on the data nature: images (keyframes and partitions) are transmitted over FTP connections and the remaining data is expressed in MPEG-7/XML [1] metadata sent over SOAP protocol

3. STUDY CASES

The defined elements aim at offering solutions to three study cases that cover the most important applications of the system.

3.1 Content ingest

When a MAM sends a new keyframe to be analyzed, this is pre-processed and added to the Content DB. The presented implementation segments the key-frame to generate an image partition and a multiscale segmentation by building a Binary Partition Tree (BPT) [2] on the top of it. Afterwards, a set of visual descriptors are extracted from each region in the BPT.

After the pre-processing, keyframes are analyzed for its automatic annotation by considering all entries in the Model DB. The detected instances, if any, are added to the Instances DB and the results of the analysis are sent back to the MAM under the form of the keywords associated to the detected instances. Additionally, the returned metadata can be enriched with a degree of confidence on the results or information related to the visual appearance and localization of the instances.

3.2 Manual annotation

Users can manually annotate regions from the keyframes through a graphical user interface (GUI), a task that generates high-confidence metadata while providing positive and negative labels for training classifiers. The annotation process begins by selecting a semantic class from the Knowledge DB or defining a new one from the same GUI. Secondly, a keyframe from the Content DB is chosen by the user and retrieved together with its partition and BPT. Finally, user marks the regions that represent an instance of the semantic class using the BPT to navigate through the keyframe regions, as described in [3].

After each manual annotation, the instances are added to the Instances DB and used to create or update a model of the annotated class. Model learning algorithms use visual descriptors from the annotated instances to train its associated classifiers. These visual features may already be among the descriptors computed during keyframe ingest and available at the Content DB. When the training process is finished, the new model must be evaluated on the whole Content DB. Due to the high computational requirements of this task, this process is only applied immediately on the previously positive annotations of the class, while the the rest of the database normally is explored later during periods of low usage of the system.

3.3 Query by region

In addition to the common query by keyword supported by most of MAMs, the presented system offers the option of querying the database with visual content and obtain a ranking of keyframes representing regions similar to the query. In this case, the user formulates its query by selecting a region with a GUI similar to the one used for manual annotation. The descriptors associated to the query region are compared with those in the Content DB and a similarity measure is generated for each pair to finally build an ordered list. If the query keyframe is not in the Content DB, it is previously ingested to generate its partition, BPT and extract its generic descriptors.

4. CONCLUSIONS

This document has presented a system architecture that satisfies the basic requirements of an automatic indexing system for video content. The proposed system works at a temporal resolution of a keyframe and at a spatial resolution of a region. The design tries to maximize the modularity in order to facilitate its final implementation in one or multiple machines as well as allowing an easy integration with an already running MAM. The current design presents two bottlenecks in terms of computation time. Firstly, during the evaluation of all Model DB entries on each new ingested keyframe; and secondly, in the re-exploration of the whole Content DB every time a model is updated. For this reason, future work will focus on techniques for indexing and exploring regions organized in hierarchical structures such as BPTs.

5. ACKNOWLEDGMENTS

This work was partially founded by the Catalan Broadcasting Corporation (CCMA) and Mediapro through the Spanish project CENIT-2007-1012 i3media, and by TEC2007-66858/TCM PROVEC project of the Spanish Government.

6. REFERENCES

[1] P. Salembier. B.S. Manjunath and E. T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[2] P. Salembier and L. Garrido. Binary partition tree as an efficient representation for image processing, segmentation and information retrieval. *IEEE Trans. on Image Processing*, 9(4):561–576, April, 2000.

[3] X. Giró, N. Camps and F. Marqués. Region-based annotation tool using partition trees. In *Poster and Demo Proceedings of the 2nd International Conference on Semantic and Digital Media Technologies, SAMT'07*, Genova, Italy, December, 5-7 2007.