Interface of Multimodal Searching on Multimedia Archives

Hildeberto Mendonça, Jean-Yves Lionel Lawson , Benoit Macq
Laboratoire de Télécommunications et Télédétection
Université catholique de Louvain, UCL
Louvain-la-Neuve, Belgium
{hildeberto.mendonca, jean-yves-lawson, benoit.macq}@uclouvain.be

Abstract—Multimedia content is very rich in terms of meanings and archiving systems have to be improved to consider such richness. This research proposes archiving improvements to extend the ways of describing contents and improves the user interaction with multimedia archiving systems beyond the traditional text typing and mouse pointing. These improvements consider a set of techniques to segment different kinds of media, a set of indexes to annotate the supported segmentation techniques and an extensible multimodal interaction to make multimedia archiving tasks more human friendly.

Keywords - multimedia archiving system;multimodal interaction; segmentation; annotation.

I. INTRODUCTION

The richness of multimedia contents makes them special resources that should be treated appropriately. Contents of pictures, videos, audios, 3D models, etc., are full of meanings, represented visually, aurally, spatially, and interpreted in many different ways. It is primordial that the evolution of multimedia archiving systems goes on the direction of rich representation of contents and extensible kinds of user interaction with the system.

Relevant works have been developed in this direction. Parageorgiou at al. [1] propose a multimedia, multilingual and multimodal research system (CIMWOS), which is robust in terms of archiving, indexation and retrieval, but limited in terms of number of supported modalities and types of contents. The work of Peng Dai at al. [2] also proposes a multimodal archiving system. However, it is used for a particular case, which is meeting scenarios, and considerable changes are needed to support other domains. Last but not least, Jyi-Shane Liu at al. [3] propose a very concise workflow with well delimited segmentation and annotation phases, but the solution only supports the annotation of pictures. Therefore, there is a lack of multipurpose multimedia archiving system that supports several media and multiple multimodal interactions.

We present a system conceived and initially developed in the context of the IRMA Project [4], which aimed to create an economically viable interface for multimodal searching and retrieving in indexed multimedia libraries for audiovisual companies, such as television channels, radio stations, surveillance companies and others. Nowadays, the system has been extended in the context of the 3D Media Project [5], adding segmentation and annotation of 3D models.

This paper initially describes a workflow that considers the standard media management in audiovisual companies and how this workflow was implemented. We go into details how the segmentation, annotation, querying and retrieving phases were designed and implemented. Finally, we list some of the relevant results and challenges that we are addressing at the moment.

II. MULTIMEDIA ARCHIVING WORKFLOW

The multimedia archiving works according to a formalized process with the global phases represented in the figure 1. The process is instantiated for each media file submitted to the system. Thus, the media file is the token that navigates through the process. The phases are:

A. Processing

The processing phase starts when the user selects a media file to add to the library. It is divided into two steps. The system: 1) calculates the checksum of the file, using the MD5 hash algorithm [6], and compares the resulting 32 bits key with the keys of existing files, avoiding duplicity. 2) extracts some values from the media to parameterize algorithms. For example, when receiving an image, the system extracts its size, resolution, colors, etc; in case of a video, the number of frames, frame rate, resolution, duration, etc.; while for an audio file, it extracts the volume, duration, energy, etc. The file will be sent to the server after the annotation or right management phase only.

B. Segmentation

At this phase, the media is available to be segmented. The segmentation consists of selecting, manually or automatically, meaningful parts of the media to describe them in the next phase. The segmentation is further explained in the following section.

C. Annotation

The annotation phase happens almost at the same time as the segmentation. For each segment created, the user has the chance to annotate it using one of the four annotation techniques. The expertise of the user may guide him/her to choose the most appropriate annotation according to the nature and complexity of the content. A section in this paper is dedicated to explain the supported annotation techniques.

D. DRM

The DRM (Digital Right Management) module defines what privileges users should have to access the multimedia content. With the availability of segmentation features, a different set of rights can be assigned, not only for the whole file content, but for each segment. It avoids, for instance, that a movie containing violence can be presented to children since they do not have access to segments that delimit violent scenes. Due to the amount of details of the DRM module and because security and right management are not the focus of this work, we will not explore this module furthermore.

E. Querying

The querying phase considers the existing annotations to find meaningful information for the user. The better annotated is the media, the more precise are the results. However, precise results also depend on queries well built, rich in terms of representativeness, and a more user friendly query interface.

F. Retrieving

The output of the querying phase is a list of media candidates to be shown to the end user. They are still candidates because a security check is still needed. The retrieving is performed in 2 steps: 1) the user receives a list of media that corresponds to the criteria of the query and also filtered with an access security check. 2) the user selects one of the media to visualize and a second security check is performed to verify if the user has rights to play the media, considering constraints such as location, age, licensing, media quality and others.

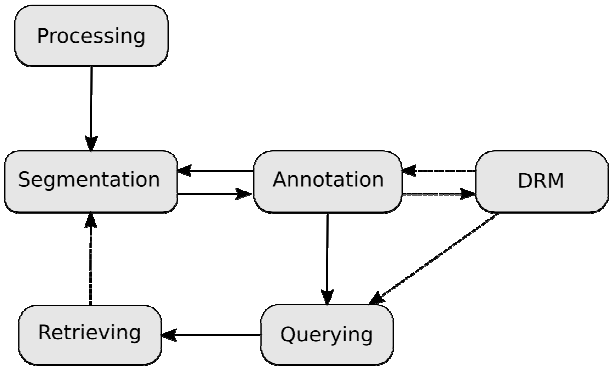

Figure 1. Workflow global view

III. MULTIMEDIA ARCHIVING ARCHITECTURE

The Multimedia Archiving Workflow (MAW) was mainly inspired by the Modality Processing Pipeline (MPP) [7], which aims to reduce the granularity of data until a level of natural understanding (from binary data to data readable by human beings). The MPP is composed of the following phases: a) detection; b) segmentation; c) meanings extraction; and d) annotation. In comparison with the MAW, this pipeline defines two different stages, the recognition and the meanings extraction, which are performed automatically to identify segments and annotations, respectively. The MAW merged these stages with the segmentation and annotation phases, respectively, because in this system, manual segmentation and annotation are also taken into consideration, while that pipeline considers an on-line processing only.

The architecture to support the MAW system was designed to provide scalability, extensibility, and robustness. It is scalable because it uses a peer-based distributed database with bi-directional replication, allowing dynamic addition of new server nodes as the demand increases. It is extensible because several existing solutions can be used with a minimal integration effort. It is robust because the chosen technologies are extensively applied on many other solutions, they have years of existence and large communities around them. The figure 2 gives an overview of the architecture.

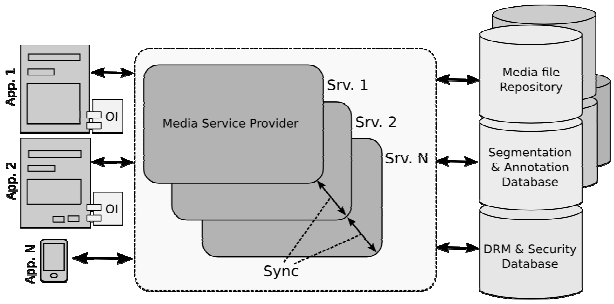

Figure 2. System architecture

The media service provider manages the information that comes from clients and organizes them in several databases. The servers assure that the file in the media repository is associated with its segmentation and annotation data and these are related with rights and access data. The media repository is a file system and references to the files are in an indexed database. Segments and annotations are stored in the document database server CouchDB [8], which supports incremental replication with bi-directional conflict detection, more appropriate for meta-data storage. DRM and security data are stored in the relational database MySQL [9]. The servers provide REST web-services [10], allowing several kinds of clients to communicate with the server, independent of programming language and platform. The role of the client side is to process heavy operations, such as the support for several modalities, automatic segmentation, automatic extraction of meanings, and also to provide rich user interaction for intuitive manual segmentation and annotation. The data is synchronized with the server, making the media and all related data available for searching and sharing.

A relevant part of the system is its support for multimodal interactions. We argue that the amount of information present in multimedia files demands enhanced user experience. Indeed, the richness of the user activity involved while interacting with such systems calls for multimodal interactions in order to provide intuitive ways of controlling the system, allowing expressiveness beyond typing and pointing.

For performance reasons, the support for multi-modalities is implemented on the client side, where the computation capacity is not impacted by the network overhead. We are using a high-fidelity prototyping solution, the OpenInterface (OI) Workbench [11], benefiting from the plethora of existing signal-processing algorithms (gesture recognition, face

tracking, sound source localization, input devices, etc.). This workbench addresses the challenge of reusing existing signal processing algorithms, ensuring seamless connections between heterogeneous native components, assuring real-time (at the interaction point of view) performance, among other benefits.

In this paper, we present a selection of ready-to-use OI components that allow us to implement multimodal interactions through the MAW. We have 4 interactive tasks implemented so far: 1) selection of segments, 2) vocal annotation, 3) query planning and 4) navigation.

IV. SEGMENTATION BASED ON RELEVANT MEANINGS

The variety of supported media demands the implementation of the following types of segmentation:

A. Spatial

A spatial segment can delimit a static region in a picture, video frame or 3D model. For pictures, videos, and 3D models, the spatial segment is bi-dimensional, containing a set of plan points involving a region of interest, S = {(x1,y1),(x2,y2),...,(xn,yn)}. The limits for x and y are based on the image resolution and the pixels gradient. For 3D models the spatial segment is tri-dimensional, S = {(x1,y1,z1),(x2,y2,z2),...,(xn,yn,zn)}.

B. Temporal

A temporal segment can delimit a sequential region in an audio, video or 3D scene. It can define when an event or information starts to happen and when it finishes, but without any spatial delimitation. A podcast, for instance, may cover several subjects and temporal segments can determine when each subject starts and ends or when the speaker says several portions of speech, or even differentiating what is speech and what is noise. In videos and 3D scenes, temporal segments can also delimit sequences, but not objects in the scene.

C. Spatio-temporal

Despite being possible to segment a unique frame of a video using spatial segment, it is rarely useful because videos might have hundreds and thousands of frames and most meanings are spread through several frames, being necessary to indicate when the meaning starts and ends. For this reason, the spatial segment was extended to consider the dimension of time, where each Spatial Segment has also a time property, and a set of Spatial Segments with the time property composes the Spatio-Temporal segment. Briefly, this is a mix of the two previous types of segment.

V. ANNOTATION OF SEGMENTS

Once the segment is created, it can be annotated using one or several techniques. The annotation phase is important to allow an efficient localization of segments based on their intrinsic meanings. A segment can be indexed using the following supported annotations types:

A. Keyword

Each keyword represents a simple word that identifies the content in the segment. Keywords are simple, efficient and widely used nowadays to create indexes of information on the web. However, it has limited representativeness when compared with other forms of annotation.

B. Transcription

It is a textual and complete description of a speech, dialog, music lyric or detailed description of a scene. Practical applications are the automatic recognition of audio sequences and optical character recognition (OCR) in images containing text. It is precise in terms of content representation, but complex in terms of computation because the extraction of meanings depends on the syntactic and semantic analysis of the transcription.

C. Domain Concepts

A good balance of representativeness and performance is the use of ontologies to annotate segments. Ontologies are used to build knowledge bases, which are composed of taxonomy of concepts from a certain domain of knowledge, the semantic relationship between these concepts, and instances of these concepts that are representations of scenarios under the modeled domain. Concepts are more representative than keywords because they can be related, but also less efficient because these relationships create a network and this network must be explored for reasoning. On the other hand, concepts are less representative than full transcription, but more efficient because only useful words are present and reasoning implies on the network exploration, and no further analysis is needed, such as syntactic and semantic ones.

VI. MULTIMODAL SUPPORT

Multimodal interactions are considered in several parts of the workflow, currently in an experimental mode to investigate the acceptance of end-users to different kinds of interactions in this specific application domain. Since we are working with association of meanings and complex contents, we believe that this kind of application demands improved representativeness of user intentions, which is sometimes difficult to represent using keyboard and mouse only.

The multimodal support is implemented in the segmentation, annotation, querying and retrieving phases. In the segmentation phase, users build geometric shapes to select spatial regions to delimit zones of interest. The interaction through hand gestures can be used to create those shapes. Thus, we implement a hand gesture input selection by combining OpenCV components (background, foreground extraction, conditional dilatation and connected components), allowing the user to select a region of interest using at least two hands [12]. In the annotation phase, users input texts and make selections of concepts. A practical way to perform these tasks would be speaking those textual values, which is especially useful when the task is performed in group, where most people do not have access to the keyboard. The appropriate modality for this is voice recognition, implemented with Sphinx-4 [13], using 8 Gaussian triphone models, trained on the Wall Street Journal Corpus, with a grammar composed of 5000 words.

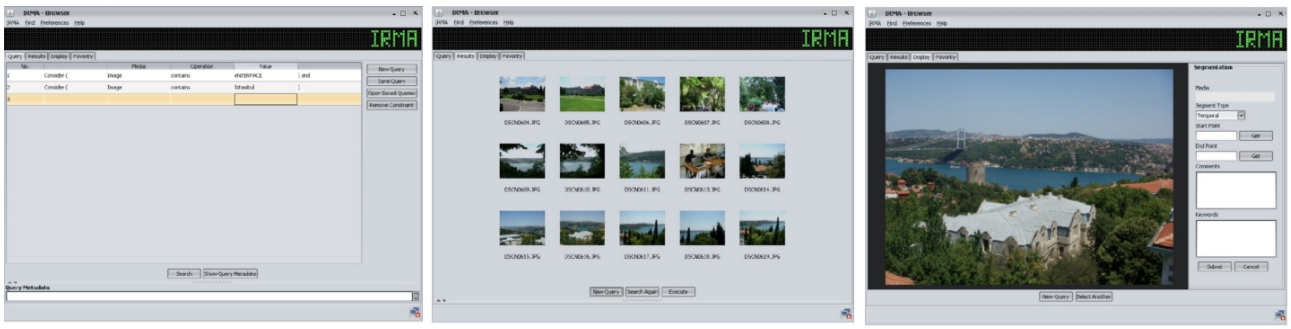

Figure 3. Figure 1 Screenshots of the querying and retrieval interfaces

The querying phase also has text input and the Sphinx-4 is also applied in this scenario. Beyond that, gestures can also be used for querying. In a typical scenario, the user describes forms (circles, squares, rectangles) using gestures and the querying mechanism interprets the user intents to find objects with that form.

Finally, the retrieving phase implements several navigation tasks, involving browsing (left, right, up, down, etc), selection and manipulation (resizing, translation, rotation, etc.). For spatial navigation, we have experienced gestures from multiple inputs including Wiimote, camera-based finger-tracking and optical flow from camera feeds. For this experiment, we have applied the "1$ recognizer" [14], a simple gesture recognition algorithm similar to Dynamic Time Warping, and a Java version of the Hidden Markov Model Toolkit (HTK) [15], depending on the type of gesture to learn. The integrated component has a 5DOF input and allows recognizing hand gestures (e.g. GloveDG5, Wiimote) and single stroke gestures (e.g. TrackIR finger tracking). A waving gesture (based on optical flow analysis) was, for instance, prototyped in order to horizontally navigate through the library.

The figure 3 depicts the querying and retrieving user interfaces, where most of multimodal interactions occurs. It starts showing the query building, the search results and the manipulation of a selected media.

VII. CONCLUSION

We have presented a multimedia archiving system that supports several media formats, three types of segmentation and three types of annotations, besides the support for multiple multimodal interactions. The system implements a formal workflow for audiovisual companies and has a scalable, extensible and robust architecture to support large datasets and reuse existing modality implementations.

The support for several segmentation and annotation techniques has been in use with good acceptance and the additional modalities, although being implemented, were not evaluated with non-experienced users. This evaluation is subject of a future work, among other aspects, such as the support of version control for media files. This is relatively challenging because in case of a new version, the content should be analyzed in order to update all existing segments since the content could be shifted, impacting on all coordinates, timestamps and positions outdated. The use of ontologies to annotate media also deserves a special attention. Different specialists modeling the same domain might produce quite different models. The development of a methodology to model domains using ontologies is needed.

REFERENCES

[1] H. Papageorgiou, P. Prokopidis, A. Protopapas, G. Carayannis, Multimedia indexing and retrieval using natural language, speech and image processing methods. Multimedia Content and the Semantic Web. John Wiley & Sons, 2005, pp. 279-297.

[2] P. Dai, L. Tao, G. Xu, Dynamic context driven human detection and tracking in meeting scenarios. VISAPP (Special Sessions) 2007, pp. 31-40.

[3] J.-S. Liu, M-H. Tseng, T-K Huang. Mediating team work for digital heritage archiving. JCDL'04, ACM, 2004, pp. 259-268.

[4] Communications and Remote Sensing Laboratory, Multimodal search interface in audiovisual content - IRMA, TELE Lab. [Online] Available: http://www.tele.ucl.ac.be/view-project.php?name=IRMA [Accessed: Jan. 25, 2010].

[5] Multitel, "3DMedia", MediaTic. [Online] Available: http://mediatic.multitel.be/platforms/3dmedia.html. [Accessed: Jan. 25, 2010].

[6] RFC 1321, The MD5 Message-Digest Algorithm. 1992.

[7] H. Mendonça, L. Lawson, O. Vybornova, B. Macq, J. Vanderdonckt. A fusion framework for multimodal interactive applications. ICMI-MLMI 2009, Cambridge, USA, 2009.

[8] Apache Foundation, CouchDB Project, [Online] Available: http://couchdb.apache.org. [Accessed: Jan 25, 2010].

[9] Oracle Corp., MySQL, [Online] Available: http://mysql.com. [Accessed: Jan 25, 2010].

[10] R. T. Fielding, Architectural styles and the design of network-based software architectures. Ph.D. dissertation, University of California, Irvine, CA, USA, 2000.

[11] L. Lawson, A. Al-Akkad, J. Vanderdonckt, B. Macq. An open source workbench for prototyping multimodal interactions based on off-the-shelf heterogeneous components. EICS '09. ACM, New York, NY, 2009. pp. 245-254.

[12] A. D. Wilson. Robust computer vision-based detection of pinching for one and two-handed gesture input. UIST'06. ACM, New York, NY, 2006. pp. 255-258.

[13] W. Walker. Sphinx-4: A fexible open source framework for speech recognition. Technical Report SMLI TR2004-0811, Sun Mycrosystems Inc., 2004.

[14] J. O. Wobbrock, A. D. Wilson, Y. Li. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. UIST'07. ACM, New York, NY, 2007. pp. 159-168.

[15] S.J. Young, S.j. Young, The HTK hidden markov model toolkit: design and philosophy. Entropic Cambridge Research Laboratory, vol. 2, 1994. pp. 2-44.