Towards a semantic multi-modalities description of audiovisual documents

Manel Fourati, Anis Jedidi, Faiez Gargouri
Laboratory MIR@CL
University of Sfax.
Sfax, Tunis
Manel.Fourati@fsegs.rmu.tn, jedidiani@gmail.com, faiez.gargouri@isimsf.rnu.tn

Abstract—The description of an audiovisual document represents a major challenge for multimedia retrieval. Indeed, the lack of descriptive metadata extraction makes these proliferated resources underused in the querying process. To overcome these difficulties, the extraction of the semantic metadata of the content and the different structures of an audiovisual document is required. In this paper, we present a method that enables an automatic description of audiovisual documents. This automatic process is mainly based on the use of a multitude of modalities for the description and modeling of audiovisual documents, the standardization of these descriptions in the MPEG-7 standard through its description definition language (DDL) and on the use of semantic web language to link between the audiovisual resources.

Keywords- Semantic Web; description; metadata; audiovisual document; MPEG 7.

I. INTRODUCTION

In the web environment, the W3C standard provides a technology set of semantic web. The use of these technologies can provide valuable assistance in the description and the search for audiovisual resources. The Semantic Web environment provides a wide range of description languages which can be useful for audiovisual resources in different films, semantic, thematic, temporal, and environmental dimensions. The objective behind the annotation is probably a better retrieval. Thus, a fairly complete and coherent description of audiovisual resources is a need in itself because of the lack of automatic processes for the description of audiovisual resources. It is in this context that our work is set. We are particularly interested, in the semantic description that represents a major challenge. We aim to define a process for the automatic description and modeling of audiovisual documents based on the extraction of metadata following a pattern of annotation. This description is based not only on its structure modeling, but also on the addition of other suitable structures representing other contexts of annotation. These contexts exploit technologies from the Semantic Web to allow a retrieval of audiovisual documents which are more relevant and more semantic. This paper is organized as follows: First, in the second section, we present a literature review of the different techniques of description and modeling of audiovisual documents. The third section is devoted to a brief overview of our general approach to the description of audiovisual documents. Finally, the last section includes the conclusions and a discussion of later works.

II. RELATED WORKS

The research on audiovisual documents is a major issue that has grown significantly in recent years. In order to facilitate access, several studies have proven that it is essential to use some organization techniques of audiovisual documents as an index [1]. Indexing is the process of linking the audiovisual content with its description. To enforce this indexing, it is necessary to extract representative descriptions of the audiovisual document and a representation of the document structure as well.

In this section, we present an overview of the most relevant works proposed in the literature related to the description and the modeling of audiovisual documents.

A. Description of audiovisual documents

The description of audiovisual documents consists in the extraction of adequate descriptions in order to obtain annotations and to generate descriptive metadata. In literature, there are three approaches to describe audiovisual documents: some works propose methods based on low-level analysis, some others are documentary directed and others are based on the audiovisual content.

1) Low-level description

The works based on the low-level analysis consider the audiovisual stream as visual data using image and signal processing tools. In [2] the authors propose an annotation method of the video visual content. This method is based on a combined approach based on Mathematical morphology operations and consistency criteria of space for text detection in video images. The technique of optical character recognition (OCR) is used to extract keywords. In [3], the authors extract descriptions as key frames of plans, like the beginning and end of a scene, by calculating the average of all the frames of the video sequence.

2) Documentary description

As part of the annotation and the description of audiovisual documents, several approaches of audiovisual documents description are based on the documentary process, namely [4] and [5]. In this context, they propose methods based on the extraction of strict documentary metadata, expressed with the MPEG-7 language. The documentary elements are then described as resources classified under the audiovisual concepts. They follow a backward description process that starts from a concept or a theory to build the necessary steps for the description. To produce descriptions, Troncy [5] articulates ontology dedicated to the audiovisual with thematic ontologies. However, M.Bui Thi [4] uses different vocabulary resources...
(dictionaries, domain ontology extracted from encyclopedias and specific dictionaries, etc.). In this context, several basic tools for the description of audiovisual documents are implemented namely, ANVIL, VideoAnnEx etc., which generate annotations following the documentary patterns (XML or MPEG7). Despite the performance proven by the documentary description works, the dissatisfaction of the user’s needs persists, because of the lack of a standard representation and the lack of semantic description. Indeed, the descriptors proposed by the standard are far from being sufficient to extract and to structure the semantics conveyed in an audiovisual document. The shortcomings of these works have been taken as a starting point for the content description modeling approaches.

3) Semantic description

In the context’s description of the audiovisual documents content, the goal is to extract meaningful and illustrative information related to parts of the audiovisual document, in other words, an instantiated meaning is associated with each fragment of the document by one or several concepts. The semantic search engines provide the data of the video’s content and represent them as key elements; one of these is Voxalead, which extracts the semantics of the speech’s content and generates an xml file containing the words identified in the audio document. The Works of [6] were devoted to the description of audiovisual documents and the development of the ADVEN platform which allow creating semantic annotations semi-automatically by the creation of hyper videos generated as XML or XHTML format. Azaiez in his work [7] used the ADVEN platform to create semantic annotations and proposed an approach of thematization of the audiovisual document according to the topics mentioned in each document in order to generate descriptive metadata.

In literature, there are some works that have shown the importance of describing and structuring the audiovisual content by focusing on filmic descriptions. In his work, P. Stockinger [8], presents different semiotic descriptions of the contents of audiovisual archives namely, textual description, pragmatic, keyword, referential, thematic, taxemic, para-textual rhetoric, narrative, audiovisual, per-textual and translation. To satisfy the need for the exploitation of Audiovisual Archives, AAR [9] mentioned the use of a tool called Interview allowing the description and the indexing to enrich an audiovisual document with annotations exported in XML format. It is annotated so that the semiotic description depends on the thematic and rhetorical analysis, and on the logical and physical organization.

In our work, we focus on extracting descriptions of audiovisual documents automatically by offering a multi-modalities documentary-oriented approach based on the audiovisual content through semiotic descriptions defined by Stokinger.

B. Modeling of audiovisual documents

In this section we present the works related to the modeling of audiovisual documents such as: the structure modeling and the content modeling of the audiovisual document. Several studies are based on the modeling of the structure of the audiovisual document. They focus essentially on the way to organize the information in the document. The level of the most axiomatic structuring is the priori segmentation of the document (sequence, scene, and plan). Előd Egyed-Zsigmond in his work [10] suggested a method for structuring audiovisual documents using layers which are interconnected by annotations (Strata-IA). It articulates a knowledge base to extract the annotations to be presented with their relationship as a graph. Further works, namely the static modeling and semantic modeling, are based on modeling the content of the audiovisual document.

Concerning the static modeling, the most conventional method of structuring in the audiovisual applications comes from the domain of databases. Some works have been devoted to the extraction of data from the audiovisual signal. [11] Suggests a data model known as OVID (Object Oriented Video Information Database) providing descriptive data for each video object. It allows sharing the common data between video objects, called "heritage, inclusion, generalization between intervals". In [12] the author used to describe and to structure the content of a video, a relational or object-oriented DBMS. However, the major drawback of these approaches comes from their dependence on the databases platforms, the heterogeneity of audiovisual resources and the flattened structure of the databases.

In the context of the semantic modeling, it is essential to link the audiovisual content to its description. In this context, [13] proposes a conceptual modeling method based on a description by concepts and conceptual relations. To provide a list of concepts, [13] must go through a step of extracting important data. To do this, the authors use a video annotation tool (video Annex) or a semantic search engine (Voxalead). However, several disadvantages are mentioned in [14] such as, the limited scalability, the concepts defined in a static way, etc. Auffert mentioned in [15], that the descriptions extracted are the knowledge about the contents of the document. Therefore, it is important to generate a set of metadata as structured information that describe the resources. In his work, Issac [16] uses the languages and tools related to the semantic web for the description of audiovisual documents. He uses a combination of ontologies represented in OWL-DL (Ontology Web Language Description Logics) and inference rules allowing a structured description and a more complete research of audiovisual sequences. The description is based on a predefined theme: medicine. The annotations are expressed in RDF (Resource Description Framework) so that all resources can be distributed and reused in other applications. He used the ontology "core" dedicated to the audiovisual with thematic ontologies. He follows a top-down approach for the description.

In our work, we focus on structuring the extracted descriptions following a structure modeling that is based on the stratification approach and a semantic oriented knowledge modeling. The remainder of this paper presents our approach of a multi-modalities description of audiovisual documents.
III. OVERVIEW OF THE APPROACH

After presenting the appropriate techniques of the description and the modeling of audiovisual documents, we will reserve this section to provide a general overview of the approach. Our approach allows us to describe and model the audiovisual documents which are characterized mainly by: the extraction of semantics from different analytical modalities, the automation of the description process and the link between audiovisual resources to improve the multimedia retrieval. Fig.1 illustrates the general principle of our approach. It consists of two phases: i) the pre-description phase and ii) the semantic description phase.

A. Pre-description phase

We are interested in describing the cinematic audiovisual documents in order to extract semantic and automatic metadata following a bottom-up gait. A multi-modalities description involves the description of multimedia documents through different descriptors from different media types. Our work is based on the following three modalities: the textual modality, the flow modality and the audiovisual modality. In the Fig.2 below, we present an example of presentation and the link between these three modalities:

Therefore we can extract generic descriptions. In [17] we presented a method that allows the extraction of a semantic and automatic of genre description from the film content. The scenario contains the discourse between the actors and the different actions of the film. By analyzing the speech, several semiotic descriptions can be extracted. The scenario is a source of information that allows us to describe at the same time the structure and the content of the audiovisual document, namely thematic and textual descriptions. Figure 2 shows an example of extracting the thematic description from the scenario. Res1 is the result of the thematic description from the scenario. This first analysis of the pre-production documents can hide some realities linked to the content of the film. That is why; we analyze the film in its post-production phase by processing the audiovisual stream.

(2) The flow modality consists of extracting the description of the content of information integrated in the audiovisual signal. This processing is based on the following modalities: the visual modality (extracting the text superposed on the images using a tool for character recognition OCR), the auditory modality (tool speech processing to obtain the text that contains the speech transcribed). By analyzing these texts, we obtain segments containing the most relevant information (Res2). After the annotation of the audiovisual document in these two modalities of pre-production and post production, we intend to complete the description with a third modality using the structure of the film.

(3) The audiovisual modality consists of extracting the segmentation of an audiovisual document. This last modality is based on the stratification of audiovisual components and aims to complete and harmonize the thematic analysis started in the previous two phases (Res4). Based on these four results, we have elaborated algorithm Annotation to identify the annotation and their adequate segments. Given that:

\[ R = \{Res1, Res2, Res3, \text{ and } Res4\} \]  

where Resi= \{Theme1, Theme2, ..., Themei\} and Segj=(Time_begin, Time_end)

Algorithm Annotation

| Input: \( R = \{Res1, Res2, Res3 \text{ and } Res4\} \) |
| Output: Segthem: Segment table of each theme |
| n: number of themes |
| k:number of segments of each Theme |

Begin
For each i in [1..n] do
For each j in [1..4] do
If (Themei in Resj) then
For each l in [1..k] do
If(Segj not in Segthem) Segthem[l]=add(segj)End if
End for
End for
End for
End

The documentation used in the descriptions phase is the base core of the automatic annotations in all three modalities of the audiovisual document. Respecting our bottom-up approach of processing, we intend to structure the annotation extracted in a hierarchical manner in order to obtain different levels of semantic abstraction. Such a
structure guarantees the passage from one annotation level to another to achieve the conveyed semantics in the document. This principle will be materialized by XML descriptors (metadata) integrated in the MPEG7 standard. In this context, MPEG7 provides a normative framework of description. It is sufficiently generic and flexible enough by its definition description language (DDL), to enrich the descriptions by adaptable structures. This structured description in XML represents a gateway between the audiovisual resource and the documentation. It will be enriched with an index allowing us to link each description to the parts or parties concerned in the audiovisual stream. Thus, the descriptions and their indexes represent the basis of the documents which can be used in the retrieval phase.

B. Semantic description phase

The results of the first phase must be considered as a starting point to develop a tool for the semantic description of audiovisual documents. The structure of the description obtained allows the creation of links between audiovisual resources. Each resource is considered as a document that has an annotation and a different index identified through their unique identifier (URI). The RDF is a data model proposed by the W3C allowing resources’ description and their relationships. By referring to the XML annotation structure (integrated in the MPEG7 standard) of the various resources, we intend to create RDF triples that build a network of semantic, textual and structural links between the different annotated resources. The setting up of this set of links is performed by the URI of the different resources. In addition, we consider increasing these descriptions with RDFS (RDF schemas). These schemas allow the definition of properties and classes of descriptors instantiated by RDF resources. On the basis of RDF schemas, we can generate other levels of schema which can be reused in the description of other resources. These steps provide multi-modalities description containing several types of links between the audiovisual resources. Once the descriptions are extracted, a semantic retrieval system of audiovisual resources can be introduced. This language must consider various modalities of description adapted to the MPEG7 standard: semantic, structural and audiovisual. Indeed, this system is based on RDF triples of description and links between RDFS schemas produced through the DDL of MPEG 7.

IV. CONCLUSION AND PERSPECTIVES

This paper presents a review of literature concerning the description and the modeling of audiovisual documents having cinematic nature. Thereafter we addressed our first reflection of the proposed approach in order to resolve the problem of description of audiovisual documents. Two phases form the backbone of our work: The first is the pre-description of audiovisual documents across different modalities of metadata and the second is devoted to the semantic description of audiovisual documents by matching a query language to the environment description MPEG7 and to the multimedia context. The bottom-up approach leads to the production of the RDF document and the RDF description schema integrated into the description environment MPEG 7 and the multimedia context. As a first perspective, we consider the validation and automation of the first phase by multi-modalities and an automatic tool for the description of audiovisual documents. As a long-term perspective, we will study the retrieval languages to choose the most appropriate language to our needs and propose a semantic retrieval process.

V. REFERENCES: