Multimedia Knowledge Representation For Automatic Annotation of Broadcast TV Archives

Maurizio Montagnuolo
Università di Torino, Dipartimento di Informatica, Corso Svizzera 185, Torino. Italy.
montagnuolo@di.unito.it

Alberto Messina
RAI Centro Ricerche e Innovazione Tecnologica (CRIT) Corso Giamone 68, Torino. Italy.

Journal of Digital Information Management

ABSTRACT*: Multimedia content classification and retrieval are indispensable tools in the current convergence of audiovisual entertainment and information media. Thanks to the development of broadband networks, every consumer will have digital video programmes available on-line as well as through the traditional distribution channels. In this scenario, since the early '90s, the most important TV broadcasters in Europe have started projects whose aim was to achieve preservation, restoration and automatic documentation of their audiovisual archives. In particular, the association of low-level multimedia features to knowledge and semantics for the purpose of automatic classification of multimedia archives is currently the target of many researchers in both academic and IT industrial communities. This paper describes our research direction, which is focusing on three points: (a) We first introduce a new taxonomy for classification of broadcast digital archives based on a novel theoretical approach. The advantage of this taxonomy is that it can provide an unambiguous representation of multimedia informative content from the relevant points of view to the broadcasters community. (b) We secondly present a multilayer multimedia database model to represent both structure and content of multimedia objects. (c) We further propose a framework architecture for building a Multimedia Fuzzy Annotation System (MFAS), and a description of our experimental plan.*

Categories and Subject Descriptors
H. 5.1[Multimedia Information System]; Data Models: **H 2.4 [Systems]** Multimedia Databases:

General Terms
Multimedia objects, Digital archives, Multimedia semantics

Keywords: Semantic multimedia retrieval and annotation, knowledge-based multimedia systems, video taxonomy

Received 10 March 2006; Reviewed and accepted 20 April 2006

Introduction

During the 20th century electronic media became very deeply ingrained in our lives, due to the rapid evolution of the Information Technology (IT) industry. In fact, the increasing power of electronic circuitry in workstations, personal computers, and consumer electronics, in conjunction with the decreasing cost of high-bandwidth and low-latency communication, made the availability of digital media content continuously increasing. With the advent of the World Wide Web (WWW), the digital TV and the global mobile communication society, users can feel a new form of perception of the world, through a wide variety of digital multimedia documents and services. Nowadays, multimedia applications are becoming very commonplace and a new media industry, which involves computers, entertainment, communication and consumer electronics companies, is arising. Telecommunication enterprises, such as telephone companies, TV and radio broadcasters, Internet Service Providers (ISP), are experimenting ways to offer new multimedia services. Hence, new customers will be attracted and new business opportunities will be created.

As large-scale multimedia collections come into view, research efforts on multimedia semantics are urgently needed, so that users can select desired contents specifying their needs at the semantic level. One important influential factor in semantic multimedia analysis and retrieval is the application domain. Smeulders et al. [1] have distinguished two application domains: the narrow domain and the wide domain. The first one "has a limited variability in all relevant aspects of its appearance", e.g. surveillance videos. The second one "has an unlimited and unpredictable variability in its appearance even for the same semantic meaning", e.g. broadcasting archives, which have the widest domain. That is, in the real TV broadcast world the same semantic meaning may characterize multiple programmes having different low-level features, e.g. a car racing and a curling match can both be classified as 'sports programmes' even if they are expressed by distinct audiovisual patterns. On the opposite, multiple semantic meanings can characterize one single programme, e.g. a football report within a newscast can be classified as 'sports content' as well as 'news content'.

Traditional content-based multimedia retrieval systems make use of low-level features such as colour, motion and texture to build access indexes to multimedia objects. However, even if the direct application of generic similarity metrics techniques to low-level features can give good results in visual matching of images, this is not suitable for identifying semantically similar classes. Thus, to reach semantic analysis and knowledge mining from multimedia content, novel techniques, such as ontology-based information retrieval, are required.

In this paper, a knowledge representation and annotation framework for multimedia is presented. The work described in the following sections is part of a wider context identified as the CAMAD project (Computer Assisted Multimedia Archive Documentation), a research collaboration involving IT industry, broadcasting and academics. This collaboration has been motivated by the fact that applications for semantic multimedia analysis, e.g. Automatic Genre Classification (AGC) [9, 10, 11, 12, 13, 14, 15, 16], are attracting growing interests from many Information Technology and Consumer Electronics industries, being suitable for practical applications for both TV companies and end-users. In fact, on one hand, TV broadcasters will be able to manage more

efficiently their multimedia archives, enabling fast video browsing / retrieval and decreasing costs for production. On the other hand, end-users will be allowed to create personalized TV programme lists using Video-on-Demand (VoD) and interactive TV (iTV) services. In this scenario an interactive receiver will perform AGC, segmenting, indexing and storing broadcasted programmes in different partitions of its integrated hard disk. Then, a user will access the desired kind of programmes or select one specified video scene simply browsing folders using a practical Graphical User Interface (GUI), which will be shown on his/her TV. Starting from this background and from the experience of previous projects and results [25, 26], the overall goal of the CAMAD project is to advance the current state of the art on multimedia semantics by applying rule-based inference systems to the automatic classification of multimedia objects. To achieve this goal, we are currently exploring how to: (a) Develop a sound and novel theoretical framework for the multimedia classification task; (b) Build a data model that fully characterizes the application domain and defines relationships between the different information layers carried by multimedia objects; (c) Intuitively represent content information about multimedia objects in order to efficiently store and retrieve them from a large collection of multimedia material. (d) Develop an automatic rule-based genre classification engine. That is, framing on the collection of information available or extractible from multimedia objects, to study and implement a system that classifies these objects, making use of inference and automatic reasoning techniques.

The remainder of the paper is organised as follows: in Section 2 we outline the system philosophy of the CAMAD framework, introducing the terminology and concepts that will be used. Since we think that the development of tools for semantic multimedia retrieval in industrial contexts has to be preceded by a preliminary analysis of the information workflow of multimedia archive production, we show in Section 2.2 a taxonomy for broadcasted TV programmes which summarises the typical metadata requirements coming from television archivists. In addition, Section 2.3 proposes a hierarchical multimedia database model that, basing on the developed data model taxonomy, collects meta, audiovisual, structural and cognitive information contained in a multimedia object. Section 3 details the design and development of our prototypical framework for building a Multimedia Fuzzy Annotation System (MFAS). Comparisons with other work in literature are discussed in Section 4. Finally, future work and conclusions are treated in Section 5.

2. A novel approach to Multimedia Classification

2.1 Extending the View: outline of a Reference Theory for the Multimedia Classification Problem

The key concept behind our approach consists in interpreting any knowledge-based system, which performs automatic reasoning with the purpose of deducting semantics from multimedia objects, as a real-world implementation approximating the behaviours of a community of intelligent sensors existing in the system and generating concepts associated to multimedia objects. As depicted in Figure 1, the central point is the concept of Multimedia Event, e.g. any spatial or temporal composition or aggregation of elementary mono-modal events. In the same theory, Multimedia Events are the entities that produce information and semantics (abstracted as *concepts*) downstream of the perception event experienced by some set of mutual-interacting intelligent sensors. Our aim is to formally develop a sound theorydefining and analysing these outlined principles. We think that such background theory allows a solid architecture for our system, in which reasoning modules, knowledge repositories and rule bases (altogether realising the implementation of the sensor community) are easily programmable and configurable to execute domain-driven tasks, as requested by the particular application environment in which we operate. Further extending the approach, the theory still holds when traditional or more modern pattern recognition machines (e.g. HMMs, SVMs) play the classification task in place of the rule-based inference modules, thus allowing for a unified approach to the problem.

2.2 A Data Model Taxonomy for Broadcast TV Programmes

Multimedia objects carry lots of information, including modalities (e.g. aural, visual and textual data), physical structures (e.g. shots, frames, audio clips), logical units (e.g.

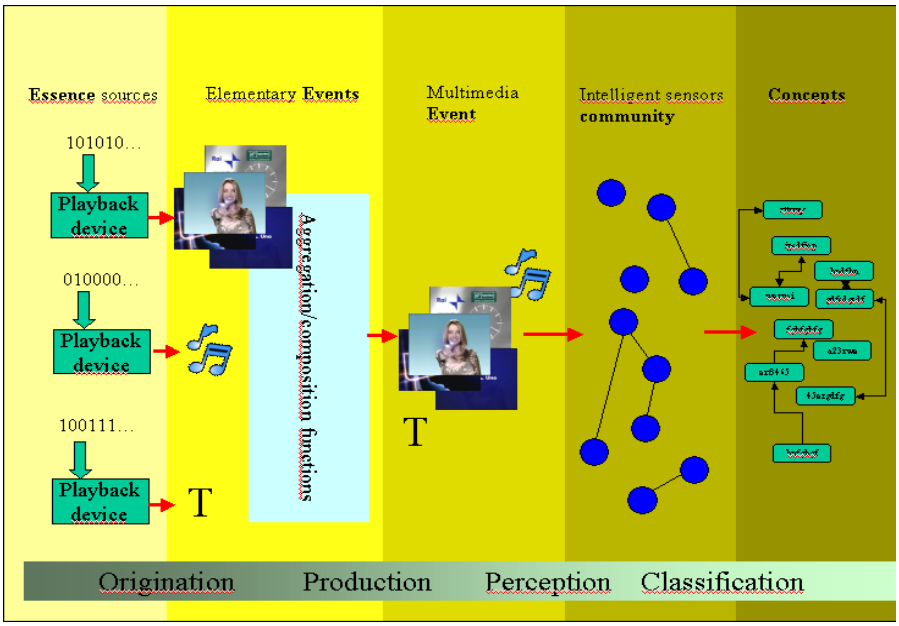

Figure 1. Workflow of intelligent multimedia generation, analysis and understanding

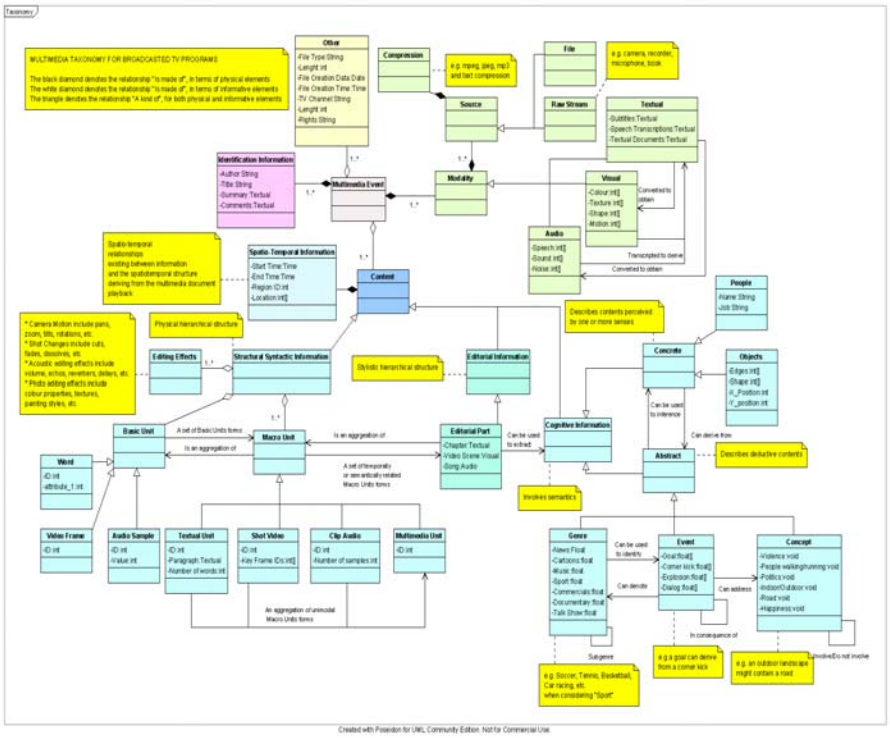

Figure 2. Data Model Taxonomy for Broadcast TV Programmes.

scenes) and semantics (e.g. events, objects, concepts). We think that an extensive use of multimedia metadata is indispensable to represent all these types of information and their relationships through a knowledge system. Nowadays, the Multimedia Content Description Interface (MPEG-7) [27] standard represents the state of the art in multimedia content description. Its major drawback is that formal semantics are not included in it. Consequently, we have developed an abstract taxonomy, which provides a technology-independent representation of multimedia content and allows to use knowledge-based techniques based on the Semantic Web and ontology-based languages [28, 29]. In the following, a UML class diagram is used to represent the proposed taxonomy for broadcasted TV programmes. Figure 2 shows the resulting model, in which a Multimedia Event is viewed as an aggregation of different parts: *Modality*, *Content*, and *Identification Information*.

Modality

The *Modality* concerns the physical properties of audiovisual objects, as they can be perceived by a multimodal sensor. Modality information is described on two layers: the perceptual layer and the source layer. The former refers to all the information that is concerned with the multimedia presentation of visual, textual and aural events (e.g. image size, colour, shape and motion characteristics). The latter refers to the characteristics of the digital sources producing the perceptual layer, e.g. *source* type (file, stream, URL), the *compression* type and other technical features. Each modality represents a particular kind of information that can be found when handling multimedia data. For instance, the visual modality bears low-level visual features such as colour, texture, shape and motion. Modalities are related to each other. For example, textual information derives from visual information using Video OCR (VOCR) applications, while Text-To-Speech (TTS) solutions permit to obtain audio data from text.

Identification Information

The Identification Information contains data that univocally identify the document like the author's name, the document title or a short summary describing the content.

Content

The Content part brings Structural-Syntactic Information, Editorial Information, Spatio-Temporal Information and Cognitive Information.

The Structural-Syntactic Information collects data inherent the spatio-temporal layout of one document. A Basic Unit is the fundamental unit of the document. It can be a Word (for text), a Video Frame (for visual streams) or an Audio Sample (for audio clips). A set of consecutive basic units forms a Macro Unit, which can be a Textual paragraph, a Shot Video, an Audio Clip or an aggregation of them (forming a Multimedia Macro Unit). Editing Effects express syntactic properties and rules between macro and basic units, and are used to separate them from each other. Editing effects are, for example, a new line to discriminate between two paragraphs in a textual document, or a dissolve to disjoin two shots in a video sequence.

Editorial Information is a stylistic hierarchical structure that provides knowledge about how the macro units are linked together to form the atomic entities of the conceptual development of a programme. For instance, a set of temporally or semantically related shots forms a scene.

Spatio-Temporal Information defines the relationships existing between the content information and the spatio-temporal structure deriving from the multimedia object playback.

Cognitive Information is related to high-level semantic concepts inferable from the fruition of audiovisual content. It is divided into two parts. First, the *Concrete* part describes contents perceived by one or more senses. Let's consider

as example *people* and *objects*. A number of applications have been developed for face and object recognition in video material. Second, the *Abstract* part describes deductive contents, such as *Genre*, *Events* and *Concepts. Genre* in the broadcast domain is typically used to aggregate documents having some common format characteristics. Its meaning includes socio-historical, cultural and subjective aspects. For instance, in video production, genre is usually intended as a description of what TV viewers expect to watch. Sub classes of genre can be associated with a main class, such as '*football', ' tennis'* and *'motor-race'* for *'sports'*. An *Event* is something that happens at a given place and time. For instance, in football sports programmes, goals, shots and corner kicks represent different kinds of events. As for genre, an event can follow and be caused by one or more previous events; for example a goal can derive from a corner kick. The majority of event detection literature deals with events in sports clips or dialogs in talk shows and movies. Genre information can be used to identify events (e.g. in a football match we are interested in finding goals) as well as from events can derive genre (e.g. explosions might infer action movies). A *Concept* is an abstract idea generalised from particular instances or occurrences, such as violence, politics, happiness, etc. Again, a concept can involve or exclude another one (e.g. a road might be present in an outdoor landscape). Concepts can also derive from events (e.g. explosions might induce the concept of violence).

Combining together different kinds of information we are in general able to derive other information, still remaining in the domain of our model. For example, given a document, its genre (a kind of abstract cognitive information), its title and its author (kinds of identification information), it is possible to infer some editorial information (e.g. because different authors and genres have different editorial styles). Spatio-temporal, cognitive and structural information can be used to optimise the compression quality and effectiveness [30] (e.g. high motion sequences and not uniform regions or faces need a lower compression ratio). Structural properties of the document (Structural Information) are usually used to automatically create tables of Content (ToC) and multimedia summaries (Spatio-Temporal Information). Finally, automatic semantic indexing can be performed considering multiple modalities and others features, such as editorial, structural and cognitive Information.

2.3 A Hierarchical and Structured Multimedia Database Model

In this section we explain our database model, which is based on the concept of *'Semantic Unit'*. A semantic Unit is a semantically closed entity identifiable during the fruition of a multimedia event. Being itself a multimedia event, a semantic unit acts as a container for descriptive data, such as the multimedia title, people/objects observed in it and categories (e.g. genre) to which the unit belongs (Identification and Cognitive Information), physical data, like the media content associated with it (Modality) and structural data, expressed both as intra-unit relationships (Structural and Spatio-Temporal Information) and inter-units interactions (Editorial Information). Applying the concept recursively, a unit results in turn a composition of other units (e.g. the RAI TRE programme "Blob" is a container of segments of other broadcasted programmes, which in that context can be considered as semantically complete entities). According to this dissertation, it is possible to model a multimedia event as a "package" containing different kinds of information, which are defined by our taxonomy explained in Section 2.2.

As it is usually possible to define a sub-event contained into a main-event, the structure discussed above could be theoretically endless. Thus, we need a model restriction that is able to express the concepts introduced above in a finite way. This model needs to support a multi-level representation that represents both structure and content of multimedia events. Our solution to the problem has been derived from the idea of partitioning the multimedia content into a set of hierarchical manageable logical units, each of them representing an independent 'semantic unit'. At the highest level there is the whole multimedia event. The subsequent levels represent others entities that are included in the previous ones, but which also exist independently (e.g. focussing on newscasts we obtain the following structure: (i) newscast programme - 'Programme Unit'; (ii) Single story - 'Sub-Programme Unit'; (iii) speech, audio tracks – 'A-atoms',; (iv) visual streams – 'V-atoms' and (v) textual parts – 'T-atoms'). Meta, audiovisual and semantic features are extracted from each unit. Thus, the information about intra-unit and inter-unit relationships is preserved. In addition related events are hierarchically linked in the database, following temporal or channel-related relationships.

3. A Multimedia Fuzzy Annotation System Architecture

This section illustrates a reference architecture for the development of a knowledge-based Multimedia Fuzzy Annotation System (MFAS) and describes the functionalities of the blocks composing the overall architecture. The proposed system makes use of fuzzy logic, knowledge representation and rule mining techniques to approximate human-like reasoning and to derive cognitive information from multimedia (e.g. genre, events, objects, etc.). Fuzzy logic techniques are preferred rather than sharp classification methods, to take into account the fact that a single multimedia object can belong to one, none or many semantic classes at the same time. Let's consider, for example, a football report within a newscast: how should we classify it? Is it 'sports' or 'news'? We think it could be both, reflecting the practices and points of view of real TV broadcast world. Rule-based approaches are chosen to model complex concepts in a simple way. In fact, rules and knowledge allow characterizing multimedia objects, basing on descriptions of their structure, low-level patterns and spatio-temporal relations. In addition, basing on acquired knowledge and past experiences, a rule-based expert system is able to update, drop, or form new rules, reflecting the natural dynamics of the television production.

A peculiarity of our system is that both the rules needed to infer high-level concepts and the semantic classes associated with them are learnt at run-time in the learning phase rather than defined a priori. Moreover, those rules and classes can also be derived from an expert's experience and knowledge [21, 22]. This way creates a rule and knowledge base on a training set selected by the expert user during the learning step. Then, using this rule-knowledge base, the system automatically assigns semantic concepts from the ontology to non-annotated multimedia sequences in the database.

In designing our system we have borrowed experience from previous RAI's projects [25, 26], focussing our attention on the scalability and efficiency of the system as well. The proposed approach results in a modular structure in which low-level features are extracted, represented and matched independently, according to different extraction / representation methods and metrics. Figure 2 depicts our MFAS, which is composed by two main modules: the ARCHISE (**ARCH**itecture for **I**nformation **S**torage and

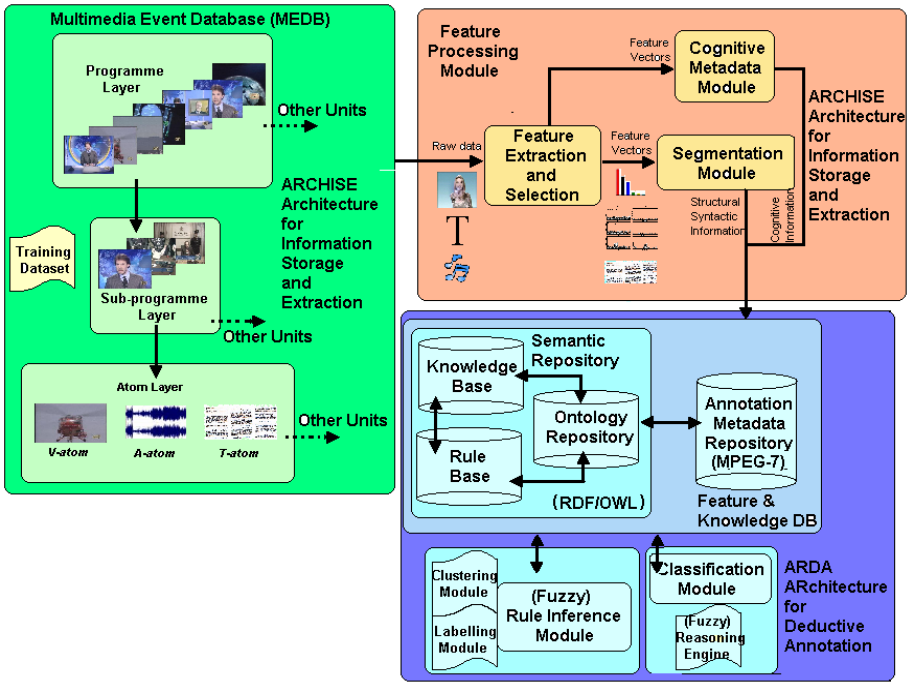

Figure 3. Multimedia Fuzzy Annotation System (MFAS) architecture

Extraction) module and the ARDA (**AR**chitecture for **D**eductive **A**nnotation) module. The former extracts all the useful information about multimedia content, according to the taxonomy presented in Section 3.2. The latter is a knowledge-based system for automatic semantic annotation of multimedia events. The following sub-sections briefly detail both the modules.

3.1 ARCHitecture for Information Storage and Extraction (ARCHISE)

The ARCHISE module is composed by the Multimedia Event Database (MEDB) and the Feature Processing Module (FPM). The MEDB is designed to store several hours of programmes from RAI broadcast channels and to store the related information in compliance with the structure described in Section 3.3, using a suitable XML document format. The FPM includes a number of available sub-systems for content information extraction (Feature Extraction and Selection Module, Segmentation Module, Cognitive Metadata Module).

The Feature Extraction and Selection Module (FESM) extracts low-level features from multimedia events in conformity with MPEG-7 descriptions [27] and stores them in the feature database. Since computational cost increases with the number of features used for classification, feature selection is needed to automatically reduce the spatio-temporal redundancy in the input data [11, 33]. Lower-dimensional spatio-temporal feature vectors are then used to compactly represent the low-level features of multimedia events.

The Segmentation Module (SM) is used to extract structural information from multimedia events. It uses a similarity algorithm based on the extracted features to perform basic unit clustering and macro unit extraction. Once macro and basic units have been detected, other audiovisual descriptors are used to enrich the set of characteristic features (e.g. average shot length, average number of detected faces, etc.). Currently, the set of low-level descriptors is composed by HSV and YUV color histogram features, two Tamura texture features [31] (contrast and directionality) and temporal and spatial activity information form. In addition, Speech-to-Text technologies are used to automatically derive text transcriptions from the spoken content.

Due to the modularity of the system, additional feature extractors will be easily added to the system (e.g. the Cognitive Metadata Module - CMM, which will include cognitive, concrete information processing tools, such as face/object detection, tracking and recognition), or disabled and enabled as desired by the particular application domain.

3.2 ARchitecture for Deductive Annotation (ARDA)

The ARDA module is basically an automatic reasoning system used to automatically infer concepts and annotate multimedia events. ARDA is composed by the following three modules: (i) the Semantic Repository (SR), (ii) the Rule Inference Engine (RIE), and (iii) the Classification Module (CM).

The Semantic Repository, which in turn includes the Ontology Repository (OR), the Knowledge Base (KB), and the Rule Base (RB), stores all information about knowledge, rules and ontologies. Ontologies are models used for knowledge representation, sharing and use. In this context, they are meant to define formal, domain dependent representations of multimedia semantic concepts (objects, genre, events, actions, etc.), their properties and the relationships among them and between them and the multimedia events. Ontologies usually make use of language terms [27, 28, 29] or, more recently, of descriptors representing also visual and aural information [20, 32]. MPEG-7 XML Schemas translations to / integration with RDF and OWL are required to derive ontological concepts based on content and structure information of multimedia events [34]. The Knowledge Base is based on the defined ontologies and contains the factual base of the domain, e.g. known multimedia data, known relationships, and any other background knowledge that an expert user may want represent in the system.

The Rule Base is a set of association rules $R = \{r_1, r_2, ..., r_n\}$, in which each rule $r : F^{m,n} \rightarrow C$

associates a low-level pattern *f'* belonging to the overall feature set *F* with high-level semantic classes belonging to the concept set *C.* ARDA uses a dataset as defined downstream of a feature selection step to generate the Rule Base during the learning phase. In addition, it takes into account all the native and derivable knowledge present in the Semantic Repository as well as the low-level and structural information of the audiovisual content contained in the feature database. A key factor that distinguishes our approach from the previous ones is that both rules and ontology concepts are not entirely predefined but also identified and labeled at runtime in the learning phase. Each time a new concept emerges the rule set and the concept set within the Semantic Repository is updated, adding new knowledge to the system. Another interesting aspect is that all emerging concepts are stored, both the ones immediately associable to well-defined semantics (e.g. *outdoors*), and the ones which are not. This is done in order to preserve in the system the maximum amount of information available from the multimedia data, and to ensure flexibility and extensibility of the classification modules. For similar reasons, input data sets are not built with respect to a predefined set of categories (e.g. *sports*, *news*, *entertainment*, *violence*), rather we expect that these concepts emerge from the analysis of data. The Rule Inference Engine is responsible of inferring rules and performs all the rule management operations, such as inferring some rules based on other ones or distinguishing meaningful from meaningless information and eliminating rules redundancy.

Once knowledge has been acquired and structured in rules, the Classification Module is used as a first proof-of-concept application to assist a human television archivist in the classification and annotation process. It uses reasoning tools to predict to which named or unnamed classes a new object is associable. Given a new Multimedia Event, firstly low-level and structural features are extracted from it. Then ARDA searches in the rule and knowledge base for finding those classes that are the closest to the Multimedia Event presented for classification. This is done by associating to it those classes showing the highest number of derived concepts (through the application of rules) and of visual and structural features (through traditional clustering techniques) in common with it. Multiple annotation strategies can be realised using a combination of different rules, ontologies and knowledge bases that the user chooses to apply at runtime among the available ones. Based on the output of the Classification Module, the annotation tools create a new annotation and store it in the Feature and Knowledge Database, thus augmenting the amount of information associable to each multimedia event and improving the knowledge of the system recursively, i.e. the augmented knowledge can be used to further refine the rule set.

4. Comparisons with other work

Research in multimedia retrieval is currently aiming at bridging the *semantic gap* existing between the physical / structural and the cognitive information of multimedia objects. As already explained in Section 2.2, physical information is concerned with the representation of the data sources providing the multimedia modalities of an object (i.e. sound, pictures and text) [2, 4], and that are used to generate the multimedia experience itself. Structural information typically pertains to data inherent to the perceivable layout of a multimedia object. It involves the representation of its spatio-temporal layout [3, 5, 6]. Cognitive information is the information pertaining to knowledge and semantics inferable from the fruition of the multimedia object. It can be derived by interpreting the structural information according to the user's experiences. Automatic semantic annotation systems aim at expressing the multimedia content in both the structural and cognitive senses, trying to compensate for the semantic gap between them. To achieve this goal, pioneer approaches firstly provided a top-down segmentation into scenes, shots and key-frames. Then this structure was linked with a Semantic Index (SI), which listed the key semantic concepts occurring in the scenes [7, 8]. Despite that, automatic and domain-independent procedures to detect shot aggregations and associate them to concepts are difficult to obtain, since yet this aggregation involves a non-trivial involvement of semantics. Thus, in the last years the research community has focused on the problem of associating multimedia objects to classes, where each class represents a specific genre or concept.

The first attempt at automatic genre classification of videos was performed by Fischer et al. [9]. Further approaches attempted to discern between few common genres, which include cartoons, news, commercials, music, sports and talk shows. These approaches made use of acoustic and/or visual low-mid level features and statistical pattern recognition algorithms, such as Hidden Markow Models (HMMs), Support Vector Machines (SVMs), k-Nearest Neighbours (k-NN), etc, achieving very good levels of success [10, 11, 12, 13, 14, 15, 16].

Analysing the solutions introduced above, it is possible to identify some restrictions that commonly affect them. First, each author has identified only few well-known categories using an approach inspired by the paradigm "It is this or that genre" (e.g. "This is sports content"). In other terms, few authors have considered the possibility that multiple genres may characterize one object. Second, spatio-temporal information between a whole programme and its sub-stories within it has been usually dismissed. In fact, typical experiments have been conducted on short clips, which have been assembled ad-hoc using semantically uncorrelated sequences. As a consequence, hierarchical and inherited relationships between the whole programme and sub-stories within it have not been taken into account in the classification models. A third problem is related to cultural and ethical aspects. Few authors have considered the fact that the television language is alive and evolves over the time, according to cultural and linguistic changes. That is, comparing clips produced at different ages or, indeed, from different TV channels, might be a difficult task, potentially introducing unneeded components in the models.

In order to overcome the restrictions presented above, more recent Knowledge-based and ontology-based approaches have been proposed. The majority of these frameworks infer classification rules from low-mid level features according to some predefined ontology of the multimedia content domain. Few examples of these systems can be found in references and many of them are specifically designed for the application domain of sports video. Zhou *et al.* [17] have classified several basketball events into different classes (*Team offense at the left/right court, fastbreak to the left/right, dunk-like in the left/right court, scoring in the left/right court* and *close-ups for audience or players*). Events have been characterized according to distinct audiovisual low-level features by a predefined semantic decision tree, achieving precision of about 78%. Jardon *et al.* [18] have proposed a rule-based approach to infer rules using fuzzy logic. In [19], an ontology infrastructure for semantic annotation of objects in car racing and football sports video is presented. Del Bimbo *et al*. [20] have developed an annotation engine that uses reasoning algorithms to automatically annotate and retrieve events in football video. Despite the fact that theseapproaches

achieve reasonable accuracy, they use predefined and static knowledge representations, thus limiting their adaptability to multi-dimensional contexts. Our approach to overcome this limitation goes towards investigating to what extent rules and classes can be inferred as a set of high-level concepts from low-mid level descriptors, and how all related multimedia content information can be expressed in a database management system. In order to obtain efficient and flexible systems, the database must model both structures and semantics of multimedia documents. Several standard description languages are suitable to express structure and semantics as well as relationships of multimedia contents, e.g. those specified in the domain of the semantic web. Some examples are the Multimedia Content Description Interface (MPEG-7) and in particular the part concerning the Multimedia Description Scheme (MDS) [27], the Resource Description Framework family of recommendations (RDF, RDFS) [28] used for encoding, exchanging and reusing structured information about web resources, and the Web Ontology Language (OWL) [29]. Based on this, we think that the observed problems can be solved using a multimedia database model that include: (a) Data representative of the physical organisation of documents; (b) Data about the cognitive content of the documents; (c) Information about all the relevant relationships between one document and other related documents; (d) Additional information, such as date of production, broadcast channel and target audience of the documents. Similar approaches have been proposed in [23, 24], in which videos are represented by a five levels hierarchy structure. The first three levels are related to the video document as a whole, while the next ones are related to its spatio-temporal segments. The first level describes the video purpose (e.g. 'Entertainment', 'Information', 'Communication'). The second level regards the video genre, such as 'Talk Show', 'Sports', 'News', etc. The third level is the video sub-genre, which involves a set of video documents that share the same genre (e.g. 'Football', 'Tennis', 'Ski Race' for sports videos). The fourth level is composed by Logical Units, which are parts of the video sharing the same concept (e.g. 'Anchorperson', 'Interview', 'Report' in newscasts). The last level includes events, which are short segments having non-zero temporal duration and invariable meaning (e.g. 'goals' in football matches or 'explosions' in action movies). This semantic-based representation of video documents is used to derive more semantically meaningful information, thus allowing for semantic-level indexing. In these trails, we have designed our taxonomy and multimedia database model, described in 2.

5. Conclusions

In this paper we have presented an overview of a novel architecture for solving the problem of multimedia classification in the application context of television broadcast archives. We plan to study and realise a prototype of this architecture in the next year. The starting point of our work is the definition of a formal background theory to provide a solid reference on which to build the development of a fuzzy rule-based multimedia annotation system. This theory envisages intelligent sensors communities as the pivotal element for the multimedia classification problem, and frames on traditional feature extraction architectures and more modern automatic reasoning techniques to achieve its goal. Then, we have already focused part of our initial work on defining a structured representation of the different kinds of information retained by multimedia events and their mutual relationships. This results in a data model taxonomy that provides a technology-independent and quite complete representation of multimedia content in the domain of television multimedia archiving independently from the practical metadata implementation. In addition, the taxonomy supplies a reference to integrate MPEG-7 metadata and the Semantic Web. According to this taxonomy, as a third step we have described requirements for developing a hierarchical and structured multimedia database that models both structure and semantic of multimedia objects. Finally, we have presented an architecture for building a Multimedia Fuzzy Annotation System that assembles the components and put in operation our research.

Summarizing, the main contribution of the CAMAD project is the development of a methodology for a tight integration of a multimedia database with knowledge management tools and ontologies to represent and annotate large broadcast TV programme collections.

Acknowledgements

The authors wish to thank Professor Maria Luisa Sapino for her help and advice in preparing this paper and Eurix Group, Turin, Italy (www.eurixgroup.com), which has sponsored the PhD of one of the authors of this paper.

References

[1] Smeulders A., Worring M., Santini S., Gupta . A.,Jain R. (2000). Content-based image retrieval at the end of the early years, *IEEE Trans. PAMI*, 22. 1349 - 1380.

[2] Del Bimbo, A. (1999). Visual Information Retrieval, San Francisco: Morgan Kaufmann Publishers, p. 270.

[3] Boreczky, J.S., Rowe L.A. (1996). Comparison of video shot boundary detection techniques, Technical Report. Computer Science Division-EECS. University of California Berkeley.

[4] Hanjalic, A., Langelaar, G.C, Van Roosmalen, P.M.B., Biemond J.,Lagendijk R.L. (2000). Image and Video Databases: Restoration, Watermarking and Retrieval, Amsterdam: Elsevier Science, p. 466.

[5] Vendrig J., Worring M. (2002). Systematic evaluation of logical story unit segmentation, *IEEE Trans. on Multimedia*, 4 (4) p. 492–499.

[6] Benini S., Xu L.Q., Leopardi R. (2005). Audio-Visual VQ Shot Clustering for Video Programs, *In*: Proc. of the 7th International Workshop DELOS AVIVDiLib'05.

[7] Naphade M., Kozintsev I., Huang T. (2002), A Factor Graph Framework for Semantic Indexing, *IEEE Trans. on CSVT*, Vol. 12(1), pp. 40–52.

[8] Li Y., Zhang T., Tretter D. (2001). An Overview of Video Abstraction Techniques, HP Laboratories Palo Alto, Technical Report HPL-2001-191.

[9] Fischer S., Lienhart R., Effelsberg W. (1995). Automatic recognition of film genres, in ACM Multimedia 1995, San Francisco, USA, p. 295-304.

[10] Truong B.A., Dorai C (2000). Automatic genre identification for content-based video categorization, *In*:Proceedings of the International Conference on Pattern Recognition, Barcelona, Spain, Vol. 4, pp. 230-233.

[11] Xu L.Q., Li Y. (2003). Video classification using spatial-temporal features and PCA, *In*: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME2003), Baltimore, MD, USA.

[12] Glasberg, R., Elazouzi K., Sikora T (2005). Cartoon-Recognition using Visual-Descriptors and a Multilayer Perceptron, WIAMIS, Montreux, May 28-31.

[13] Dimitrova N, Agnihotri L., Wei G. (2000). Video classification based on HMM using text and faces, In European Signal Processing Conference, Tampere, Finland.

[14] Liu Z., Huang J., Wang Y. (1998). Classification of TV programs based on audio information using hidden Markov Model, *In*: Proc. of the IEEE Signal Processing Society Workshop on Multimedia Signal Processing.

[15] Roach, M.J., Mason J.S.D., Pawlewski M (2001).Video genre classification using dynamics,*In:* ICASSP'2001.

[16] Dinh, P.Q., Dorai C., Venkatesh S. (2002). Video genre categorization using audio wavelet coefficients, *In:* ACCV 2002.

[17] Zhou W. , Son Dao,Jay Kuo C.C. (2002). On-line knowledge- and rule-based video classification system for video indexing and dissemination, Information Systems, Vol. 27 (8) p. 559-586.

[18] Jardon, R.S., Chaudhury S., Biswas K.K. (2002). Generic Video Classification: An Evolutionary Learning Based Fuzzy Theoretic Approach, *In*: Proc. of Indian Conference on Computer Vision Graphics and Image Processing.

[19] Dasiopoulou S., Papastathis V.K., Mezaris V., Kompatsiaris I. and Strintzis M.G. (2004). An Ontology Framework For Knowledge-Assisted Semantic Video Analysis and Annotation, *In*: Proc. of SemAnnot 2004, Hiroshima, Japan.

[20] Bertini M., Del Bimbo A., Torniai C. (2005). Enhanced Ontologies for Video Annotation and Retrieval, ACM MIR (Multimedia Information Retrieval) Workshop, Singapore, November 10-11.

[21] Dorado, A., Calic J., Izquierdo, E. (2004). A rule-based video annotation system, *IEEE Trans. on Circuits and Systems for Video Tech*., 14 (5). 622-633

[22] Wang D.H., Ma, X.H. (2005). Multimedia data mining for building rule-based image retrieval systems, *IEEE International Conference on Multimedia,* & Expo, pp. 197-200.

[23] Roach, M., Mason J., Xu L., Stentiford F. (2002). Recent trends in video analysis: A taxonomy of video classification problems, Image Processing, Video Streaming and Broadcasting, IMSA 2002.

[24] Snoek C.G. and Worring M. (2003), Multimodal video indexing: A review of the state-of-the-art, In Proc. Multimedia Tools and Applications.

[25] Del Pero R., Dimino G., Stroppiana M. (1999). Multimedia Catalogue – The RAI Experience, EBU Technical Review n° 280.

[26] Messina A., Airola D (2005). Automatic Archive Documentation Based on Content Analysis, IBC2005.

[27] ISO/IEC 15398 Multimedia Content Description Interface, 2001.

[28] WWW Consortium (W3C) (1999). Resource Description Framework (RDF), http://www.w3.org/RDF/

[29] WWW Consortium (W3C) (2004). Web Ontology Language (OWL),. http://www.w3.org/2004/OWL/

[30] Wang Y, Ostermann J., Zhang Y.Q (2001), Video Processing and Communications, Prantice Hall, pp. 595.

[31] Tamura, H., Mori S., Yamawaki, T. (1978). Texture features corresponding to visual perception, IEEE Trans. on Systems Man Cybernet, 8 (6) 460-473.

[32] Bloehdorn S. et al., (2004). Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning, *In:* Proc. of EWIMT the Workshop.

[33] Sahouria E., Zakhor, A., (1999). Content analysis of video using principal components, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9 (8) 1290-1298.

[34] WWW Consortium (W3C) (2006). Image Annotation on the Semantic Web: Vocabularies Overview, http://www.w3.org/2001/sw/BestPractices/MM/resources/Vocabularies.html

Born in 1975, **Maurizio Montagnuolo** received his Laurea degree in Telecommunications Engineering from the Polytechnic of Turin in 2004, after developing his thesis at the RAI Research Centre. Currently, he is attending the Ph.D. course in "Business and Management" at the University of Turin, in collaboration with RAI, and supported by EuriX S.r.l., Turin. His main research interests concern the semantic classification of audiovisual content.