AI-STRATA : A User-centered Model for Content-based Description and Retrieval of Audiovisual Sequences*

Yannick Prié[1], Alain Mille[2], and Jean-Marie Pinon[1]

[1] LISI 502, INSA Lyon, F-69621 Villeurbanne Cedex, France
Yannick.Prie@insa-lyon.fr
pinon@if.insa-lyon.fr
[2] LISA, CPE-LYON, F-69616 Villeurbanne Cedex, France
am@cpe.fr

Abstract. We first insist on the need for conceptual and knowledge-based audiovisual (AV) models in AV and multimedia information retrieval systems. We then propose several criteria for characterizing audiovisual representation approaches, and present a new approach for modeling and structuring AV documents with Annotations Interconnected Strata (AI-STRATA). This consists in analyzing AV documents through analysis dimensions allowing the detection of objects of interest of any type (structural, conceptual,...). Annotations are structured by annotation elements (AE) representing both objects of interest and relationships. A knowledge base is used in order to monitor the annotation process. We show how to use annotations to link different strata on the base of explicit or implicit contexts and how AI-Strata can be used to build contextual views of a stratum, using both annotation and knowledge levels. We finally show how the model can efficiently support different description tasks such as indexing, searching and browsing audiovisual material.

1 Introduction

As computer and network capabilities grow, so do the size and the number of digital data repositories, while the notion of document evolves to include new technical developments. Terminologically speaking, a multimedia document is likely to become the standard, while mono-media documents may be considered as restrictive. So, there is an urgent need for multimedia information retrieval systems capable of dealing with new digital media like images or videos. In this emerging field of multimedia information management, we will focus on AudioVisual (AV) information systems.

After a short survey of recent trends in audiovisual information retrieval systems, we will present SESAME[1], a project with a user-centered and description-

* This work is partially supported by France Télécom (through CNET/CCETT), research contract N° 96 ME 17.

[1] Multimedia and Audiovisual Sequences Exploration enriched by Experience System.

unified approach. We will then concentrate on audiovisual modeling and detail AI-STRATA, the SESAME modeling approach, before studying what services could be provided considering the fundamental task of description.

2 Audiovisual Information Retrieval Systems

Research in audiovisual information retrieval systems has greatly increased in the last decade, mainly among the image processing and the database communities. Many efforts have been done to compute image features in order to build tools using them as retrieval medium [3], or to propose audiovisual extensions to databases [17] [2]. As multimedia data do not fit exactly into classical database schemes, researchers are getting aware that tools for managing and organizing visual information could take advantage of using concepts and algorithms issued from other domains including information retrieval and artificial intelligence [13] [11]. These techniques should allow problems of similarity querying and visual browsing to be dealt with. Text-oriented query systems disappear behind environments allowing to visualize the "content" of multimedia documents, to browse visual objects, and to visually query or interpret results[2].

At the same time, guided both by this new awareness of the fundamental characteristics of visual data and by a current trend in information retrieval [4] (development of graphical interfaces and generalization of browsing to IR) novel approaches are focusing on *(re-)inserting the user in the heart of the system* for a real cooperation between human and machine. Relevance feedback [16] and machine learning are applied to similarity queries. The way people search a visual database is studied [12] and navigation through semantic ontologies of visual information are considered [5].

Studying current information retrieval systems leads to the observation than one has to *cooperate* with an audiovisual information retrieval system for performing the following tasks:

- *Indexing* is the task of describing a new document according to a model in order to organize its insertion in an index.
- *Querying* is the task of designing a query in order to find out what in the base could match it. A query can be very sharp or vague, or be an example.
- *Analyzing* consists in describing an AV document extensively in order to detect any regularities or known structural forms in AV documents concerning montage, stories, characters, camerawork, *etc.*
- *Browsing* can be seen as very precise querying (from the machine point of view), but also as wandering through the database (from the user point of view).
- *Editing* audiovisual documents is a task that can be performed to modify existing documents, to create new ones, but also to create visual summaries or clips [14] [6] guided by a set of selected descriptors.

[2] Though audiovisual systems are both concerned with visual and aural modalities, research has up to now mainly focused on the visual one. Audio information are nevertheless more and more being used.

We consider that all these tasks can be thought of as "describing a piece of video for...", which implies that the *description* of audiovisual documents should become a central task in a well-designed AV information retrieval system. Therefore AV representation (in the machine) and AV presentation (to the user) should be very close conceptually, so that user-centered visual interface could support real user-machine cooperation for description of AV documents.

Audiovisual representation should take into account every available feature provided by audio and video processing techniques. However this cannot be enough for sufficient and efficient description and representation of the audiovisual content. Higher level conceptual modeling should be used to organize this material, and a knowledge approach appears to be a necessity, either for itself [10] [22] or in order to reduce the search space before launching low-level features searches [5]. Current work on the future standard MPEG7 [18] focuses on this knowledge-based approach for conceptual modeling.

3 SESAME

SESAME is a project supported by the CNET[3] to propose and study a global approach to exploit the potentially huge repositories of audiovisual documents. The aim of the project is to take into account all facets of the problem and several research teams of different French laboratories[4] are involved in its development. Industrial partners such as INA[5] and FRANCE 3[6] are directly associated. First of all, a global architecture has been defined as sketched in figure 1. Audiovisual chunks stored in repositories are reached through High Speed Networks[7] (HSN). Audiovisual chunks are indexed or accessed on different types of clients: Indexing Clients (IC) and Accessing Client (AC), and are managed on servers by an Audiovisual Documents Base Management System (ADBMS) and a Parallel Access Engine (PAE).

The present paper focuses on the modeling approach proposed to describe audiovisual chunks in such an architecture.

The model aims to define basic elements of a content-oriented describing process. The level of granularity of an audiovisual chunk varies indeed from a whole document as it was produced originally to a small piece of an audiovisual stream focusing on a particular object. Moreover, as argued above, users accessing audiovisual servers have different needs and strategies depending on the task they want to achieve. A description model of audiovisual chunks has to consider such various levels of interest and has to be flexible enough to be used in a variety of tasks. Such a model must offer:

[3] CNET: France Télécom Center for Research and Development.
[4] LIP-ENS-Lyon, LISA-CPE-Lyon, LISI-INSA-Lyon and RFV-INSA-Lyon.
[5] INA: French TV and Radio Archives.
[6] FRANCE 3 is a french public T.V. channel.
[7] Local area networks and/or world area networks.

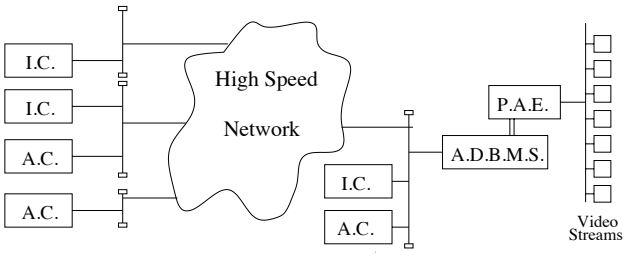

Fig. 1. *Sesame global architecture*

- a chunking process based on primary annotation and semantic annotations at any level of abstraction (from computed characteristics of images to complex concepts of the real world),
- canonical and end-user points of view to fit different goals and tasks,
- relationship links between annotations in order to build dynamic contexts as efficient user-centered filters.

The AI-STRATA model aims to offer such possibilities through task-oriented man machine interfaces, where several assistants could help the describing process. SESAME is concerned directly with image processing assistance to automatically or interactively annotate audiovisual chunks, whereas a case-based approach would help to design "winning" queries on the servers (primary indexed audiovisual chunks) from a local search description (secondary indexed audiovisual chunks).

4 Modeling Approaches of Audiovisual Documents

Audiovisual is both a visual and accoustic *sequential* medium, with a strong part of fixed *temporality* (unlike text) leading the viewer to the illusion of reality, and, by using montage, to the possibility of telling stories, with contextual relations between parts of documents.

4.1 Criteria for the Comparison of Modeling Approaches

The fundamental concept concerning a modeling approach for sequential media is *annotation*, which consists in attaching an annotation (a description) to a piece of the considered AV document. Each piece is bounded by two instants t_1 and t_2 separating it from the starting instant t_0 of the stream.

Main criteria to analyze the different approaches to describe contents of an audiovisual stream are:

- the *time granularity* of the representation model : from a time interval corresponding to a frame to a whole document (*e.g.* a two hours movie), everything is possible. *Cutting* can also be regular or not, and possibly multi-layered (*e.g.* the document is annotated as a whole plus its shots plus its sequences [7] [22]).

– the *complexity* of the annotation. This can be *simple* (a term, [6]) or *sophisticated* (for instance an iconic phrase [10], or a full description with a spatio-temporal logic [7]).
– the kind of *characteristics* represented by the annotations. We propose to use the term "characteristic" both for *primitive features* (automatically extracted from the stream: histograms, shapes, camerawork) and for higher conceptual characteristics, semi-automatically or manually extracted from the stream and from other documents (characters, actions, actors, comments).
– the *structure* of the document. As the granularity level goes thiner as one goes deeper in the document it becomes necessary to link up the different pieces that have been annotated. The resulting representation of the document can therefore be *implicitly* structured, as a result from time continuity (shots follow one another); or *explicitly* if a global structure is set up, which will be most of the time hierarchical (shot/sequence/document).

4.2 Two Main Structural Approaches for Annotation

Two main structural approaches arise from the litterature: *segmentation* and *stratification*. *Segmenting* an AV document consists in cutting it up into predefined pieces (*e.g.* shots), which will be annotated later. A structural organization can be set up to make explicit relations between pieces (as time granularity often corresponds to structural decomposition). This approach is most often used, and the present trend is to consider a three layers structure (shot/sequence /document, see figure 2), a structured annotation (usually records of attributes/ values) being associated with each piece of document (see Corridoni & al. and their "filmic grammar" [7] or [22]).

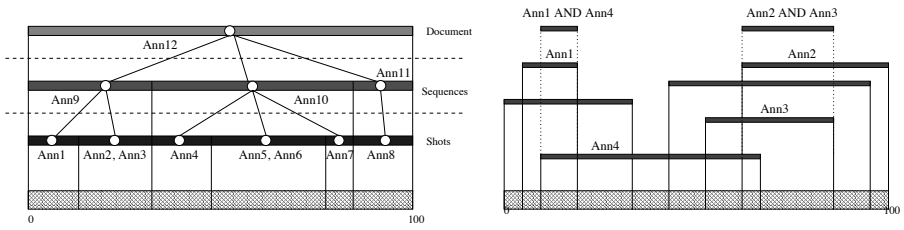

Fig. 2. *On the left, the segmentation approach: describing predefined AV pieces possibly organized in a structural hierarchy — On the left, the stratification approach: annotating (hence defining) pieces of AV documents, before* a posteriori *"useful" cutting up.*

The *stratification* approach [9] [10] differs from the segmentation one in several ways: in the segmentation approach, shots or sequences just represent themselves, the montage is forgotten, and there is a loss in the continuity of the AV media. Oppositely, a *stratum* can be any piece of an AV document to which an annotation is attached, mainly atomic, possibly expressing icon sentences.

Queries result in "useful" cuttings of the document, through strata intersection (see figure 2). No structure-oriented annotation is considered.

The essential difference seems to lie in the definition of the temporally situated and annotated pieces of documents. In segmentation, the document is first tiled with generic tiles like shots and sequences, and so cutting exists before annotation. The pieces are annotated later (classical paradigm in document description). Oppositely, in stratification they are temporally situated annotations that create the strata, while other pieces (eventually structured) emerge *a posteriori* from queries[8]. There are in fact two different cuttings of the AV documents, *a priori* and *a posteriori*, reflecting two different status for the annotation.

We think that although the stratification approach is useful for taking into account AV temporality and does really fit it, it unfortunately does not allow to consider relations between strata, for instance well known structural data though considered in recent segmentation models.

As a conclusion, this study shows first that the *dynamic* aspect related to a temporal stream should be taken into account, which implies the use of *strata* and the *atomicity* of the annotations. Second, it seems necessary to structure the annotation in order to increase the expressive power of the description, at any level of complexity. Third, contextual relations in audiovisual documents have to be considered both explicitly and implicitly.

5 AI-STRATA

The general principle of our approach is simple: we consider everything that can be revealed or said about a video, features, high-level characteristics, structural notations as characteristics temporally situated in the stream, as terms in relationship with strata. Then, in order to increase the expressive power of the representation, we set up inter- and intra-strata relations between annotations, *in the same way as annotating*. Lastly we really consider any annotation (and its relations) as *a support for contextual relations* that allow and guide contextual annotation for supplementing primary annotation.

5.1 Definitions

We call *audiovisual objects of interest* the entities that can be spotted when watching/listening to an AV stream. They can be considered as the conceptual and cognitive facets of AV characteristics. An object of interest can refer to any AV feature, from a low-level abstraction (a color histogram) to a high-level one (an action). There are as many objects of interest as there are possible analyses, so we group into *analysis dimensions* the analyses that allow to spot the same kind of objects of interest. We can for example consider *analysis dimensions* linked with shots, faces, activities, people detection, or more specialized

[8] Of any type: looking at two time lines and the associated annotations on Mediastream [10] is a kind of "visual" query.

objects, like "President Clinton", or more general ones, like structural unit spotting. Analysis dimensions are just a way to group detection methods or types of characteristics in a relevant manner regarding the goal of the subjacent analysis.

An *audiovisual stream* is no more than a file with audio and video data, beginning at t_0. We define an *audiovisual unit* (AVU) as an abstract entity representing any stratum of the stream. An AVU is created whenever its existence becomes of interest, that is when it has been spotted as a stratum linked with an object of interest: "this is a part of an AV document called a shot", "this is a part of an AV document in which appears X", *etc.* Every AVU must, by definition, be annotated, associated with a characteristic from the spotting of which it has been established: this is the primitive annotation (see figure 3).

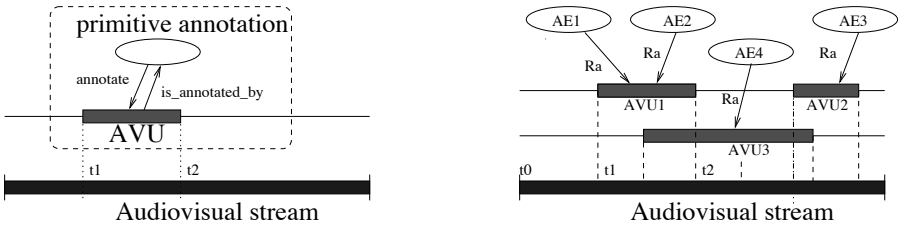

Fig. 3. *On the left: one audiovisual unit and the primitive annotation that creates it — On the right: audiovisual units and annotation elements*

We call an *annotation element* (AE) a *term* in *annotation relation* with the AV unit, as the symbolic expression of a characteristic. As analysis can spot many types of objects of interest, there are many annotation elements associated with color analysis, shapes, camera movements, audio features, objects, activities, types of documents, sensations, *etc.* For instance $\langle Shot \rangle$, $\langle Clinton \rangle$, $\langle Round\_object \rangle$, $\langle Zoom \rangle$, $\langle INCLUDED\_IN \rangle$, $\langle Music \rangle$ or $\langle Sad \rangle$. The fundamental principle lies here in the *atomicity* of the AE: it is possible to express any characteristic as an AE. AEs can have attributes for numerical expression of features (a color histogram will appear as an attribute of an AE indicating that this histogram was computed). Other AE attributes can also be added: texts, images, sets of features for similarity search, *etc.*

To complete the primitive annotation that defines an AVU, it is possible to add as many AEs as necessary to annotate it (*cf.* figure 3), in two differents ways. First, by *grouping* annotations of the same temporal range: we can add to the annotation element $\langle Document \rangle$ (annotating an AVU corresponding to a whole AV document) other AEs regarding for instance the author or the producer. Second, by *structuring* the annotation using annotation elements (see part 5.2). We call *direct annotation* from an AVU the set of AE in annotation relation R_a with an AVU.

5.2 Relations Between Annotations

We have seen earlier how it was important to structure the annotation to be able to express more complex information pieces of AV documents. To achieve this goal it is necessary to express relations between atomic annotations we have already set. If some relations can be considered as implicit (two AEs, corresponding to two people, annotating the same AVU probably shows that the two characters appear together in the video) while other ones have to be explicit. For instance "this shot is included in this sequence", "this character has that activity", "this object, the sun, is linked with that round, yellow form", "this shot is re-used in that document", and so on.

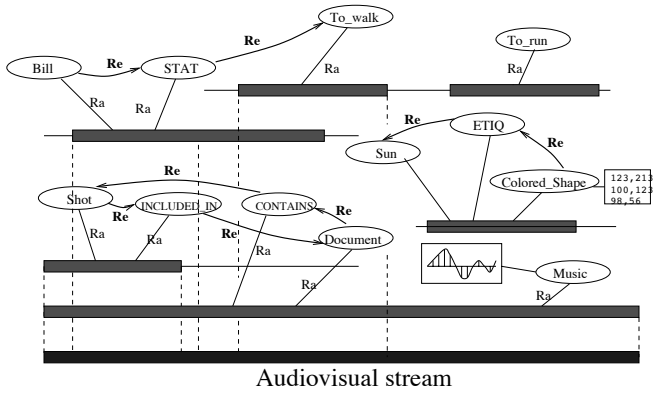

Fig. 4. *Structuring the annotation.*

This can be done by keeping the stratification principle, using the atomic annotations already set, *in the way* we have already annotated. We define one single relation between AEs (the *elementary relation* R_e), and choose to *name* relations that matter as annotation elements. For instance, to express the fact that "Bill is walking" starting from two AEs $\langle Bill \rangle$ and $\langle To\_walk \rangle$[9] annotating two different AVUs, we connect them by a "canonical" annotation element $\langle STAT \rangle$[10] which is used also to annotate the first AVU. We set up two elementary relations to express both the fact that "Bill's walking" (see figure 4). Considering a relation like "this shot is included in this sequence", we should use the canonical AEs $\langle INCLUDED\_IN \rangle$ and $\langle CONTAINS \rangle$.

Finally, annotating a stream consists in studying it through as many analysis dimensions as necessary, so that AEs and the associated AVUs emerge. Relations between AEs are then set up with canonical relation AEs and linked up by elementary relations, as illustrated by figure 5.

[9] Spotted along two analysis dimensions corresponding for instance to "characters" and "activities".

[10] Inspired by Sowa's Conceptual Graphs [20].

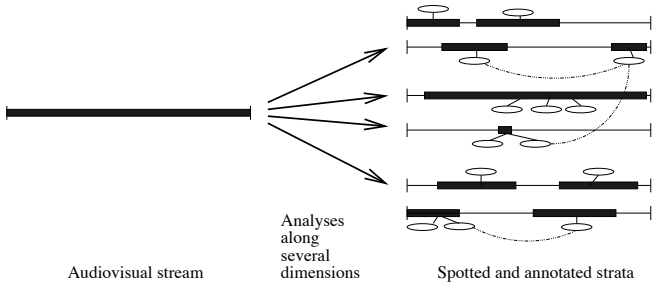

Fig. 5. *The annotation process: from raw AV material to annotated and linked strata*

5.3 Knowledge Base Associated with the Annotation

Annotation elements are supports for searching contextual relations, for infering from the network of annotations that constitutes an AV document (see figure 8) annotated by, for instance, an archivist from a broadcasting company. In order to facilitate and monitor later searches, it is necessary to consider AEs as terms issued from a vocabulary controlled within a knowledge base: an AE is instanciated from an *Abstract Annotation Element* (AAE) by the way *inscription in the stream*. An AAE is in *stream inscribing relation* R_{si} with one or several AE while an AE is in *decontextualizing relation* R_d with one AAE.

Abstract annotation elements are organized in a knowledge base with concept-relations (for instance specialization, equivalence, part-whole,...) between AAE-nodes, see figure 6. The knowledge base as an ontology appears like a thesaurus in classical information retrieval, but, as it is in fact an ontology, it should also be able to organize knowledge about possible AE attributes (value domains, by default or privileged relations between AE).

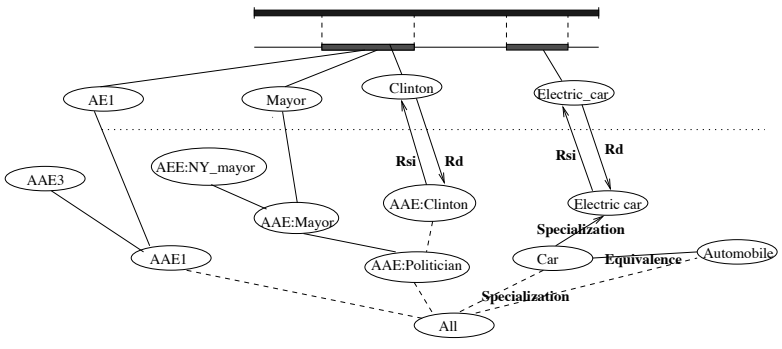

Fig. 6. *Abstract annotation elements and knowledge base*

The description of an AV document leads to a set of AVUs and AEs. This set is not *a priori* independant from the system, it is a part of it. It merges into the *Annotated Audiovisual Document Base* (AADB — set of all the AVUs and AEs of the system) taking into account the relations that some AEs have with other AEs from the base. The AV stream/file does not only appear *per se* in the base, but also as a set of annotated AVU in the AADB, linked with the knowledge base (see figure 7).

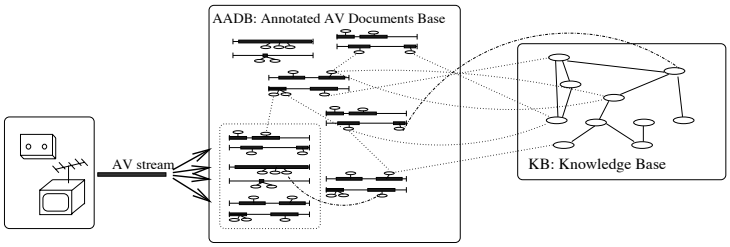

Fig. 7. *Stream annotation and system bases*

5.4 Context from an AVU and Construction of Views

We insisted above on the importance of context in the audiovisual domain, at the intra-document level: shot context ("this music goes together with this character"), montage context ("this shot is only understandable if preceeded by this") ; but also at the inter-document level ("this shot has been extracted from this document").

As it is possible to set up relationships between AEs annotating two different AVUs, we should be able to consider different AVUs in relation to each other, one being part of the context of the other. This context can either be a *temporal* one (one AVU has a temporal relation with all the AVU of the stream), a *structural* one (two AVU annotated by the AE $\langle Sequence \rangle$ and $\langle Document \rangle$) or — and this is the most general case — a *conceptual* one.

For example if many AEs corresponding to a given character are inscriptions (*i.e.* instances) in the stream of the same AAE, or if there is an explicit relation between the AEs $\langle Zoom\_in \rangle$ and $\langle Car \rangle$ specifying that "the zoom-in of the camera focuses on the car".

We consider that every contextual relation we have evoked can be thought of as a relationship between AE annotating AVU spotted along well-choosen analysis dimensions. In other words, we think of *annotation elements* not only as simple annotations, but also as starting *elements* for more complex annotations in their net organization, and eventually, out of this organization, as *media* for contextual relations.

We define two contexts for an AVU:

– the context *internal to the annotated documents base*, composed of AVU related to each other with an *internal contextual relation* $e_1 R_e \ldots R_e e_2$ between some of the AE annotating them.
– the context *external to the annotated documents base* concerns a set of abstract annotation elements (AAE) in *external contextual relation* with annotation elements from the AVU.

In figure 8, *avu*2 and *avu*4 are in the internal context of the AVU u, while AAE11 and AAE12 belong to its external one.

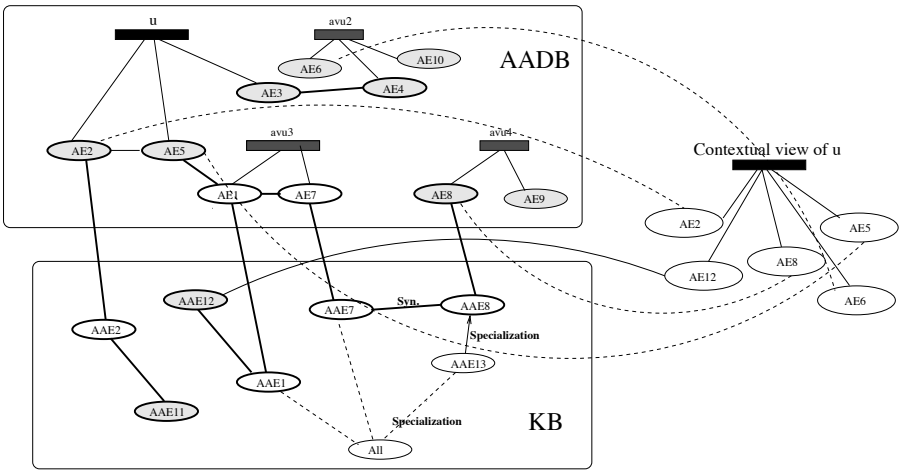

Fig. 8. *First step in view construction: building the context of an AVU ; second step: choosing the contextual AE*

The notion of context so defined, we now consider how we can build *contextual views* of AVUs, made up with *contextual annotation elements* coming from their context. These AEs can come from the internal context (AEs formerly annotating another AVU) or from the external one (inscriptions in the stream from AAEs).

A contextual view is constructed in two steps: the first one consists in constructing the context of the AVU and determining what AEs are candidates for contextual annotation: in grey in figure 8 we have AE2, AE5, AE3 (u is in its own context); AE4, AE6, AE10 (from *avu*2); AE8, AE9 (from *avu*4); AAE11 and AAE12. The second step consists in choosing the ones among the potential contextual AE which will actually belong to the view of the AVU: some issuing from the internal context will already exist (fig. 8: AE2, AE5, AE6, AE8), while others (coming from the external context) will be inscriptions in the stream of AAE (see AE12).

These two steps in the construction of a contextual view are controlled by two filters called the *contextual filter* and the *descriptive filter* together representing the *view filter*. We see that the simplest view filter would associate to an AVU

the set of the AE annotating it. This is called the minimal view, as the filtered context would be minimal (the AVU itself) and the filtered description would be maximal (all the potential AE are selected).

View filters are defined according to what users want to do: for a large search in the context of an AVU, the scope of the context should be limited in depth by a maximum number of elementary relations between the elements of the context and the AVU. However, more specific search should narrow the context (for instance, only the context linked with camera movements and people) and the choice of particular potential AE (only politicians, or actions...). Using view filters allows any kind of describing task to fit with particular needs and goals. As the whole system (AVUs+AEs+AAE's) appears much like a graph, tools and algorithms that will be designed and used to manage it will be strongly related with graph theory and practice: graph manipulation and subgraph isomorphisms will indeed play an important role in AI-Strata management.

5.5 Discussion and (closely) Related Work

We have presented our approach for AV modeling with Annotation Interconnected Strata, which consists in considering every possible "comment" about an AV document as an object of interest spotted along an analysis dimension. The object is represented by an annotation element defining an audiovisual unit. Annotations are structured thanks to relations between atomic annotations (hence between strata). This structuring is made by annotating. The vocabulary of the terms-AE is controlled with a knowledge base organized as a semantic net. The "deconstruction" of a document resulting from its annotation allows the explanation of every contextual relation inside the document (even between AVUs that do not temporally meet, which goes further than the simple juxtaposition of annotations), but also with other documents in the base. Then it becomes possible for any AVU to consider a context based on the relations its AEs have with other AEs or AAEs of the Knowledge Base. Annotations act as a medium for establishing conceptual contextual relations (among which temporal and structural relations take place), allowing propagation of annotations.

In [23] is presented a video description model called "time-stamped authoring graph", where textual annotations are attached to time instants. Annotations are then connected using three types of links: commonsense, generalization and normal links. Retrieval uses keyword match and time interval calculation with time-stamps. This approach is related to ours as it also leads to an annotation graph for AV documents, but it differs in at least two important features. First, no controlled vocabulary nor conceptual relations are available (for instance generalization only occurs in the context of a document, and is not a knowledge link). Second, the annotations are time-stamped, while we on the contrary introduced the concept of audiovisual unit, which acts as a mediation between the audiovisual stream and the annotation, with no intrinsic semantics[11]. Using this

[11] The meaning of the AVU (hence of the AV piece of the AV document it represents) is provided by the AVU annotation (*i.e.* its set of annotation elements), but also by the

neutral and intermediate level between AV data and annotations could bring an elegant solution to the conflict between segmentation and stratification stressed ealier (see table 1).

	Segmentation	Stratification	AI-STRATA
Time-granularity	linked to hierarchical struct doc/seq./shots	not limited	not limited
Complexity	records Att:Val	single icon or icon sentence	not limited (context)
Characteristics	not limited	issued from conceptual categories	not limited
Structure	structural tree hierarchy	no structuration	not limited

Table 1. *Comparison between segmentation, stratification and AI-STRATA approaches*

Second remark: it is not mandatory to use temporal relations (co-occurrence [10], "interval-inclusion based inheritance" [17]) or structural relations (a shot can inherit annotations from a document it is a part of [21], a sequence is enriched by annotations of the shots that are part of it [6]) relations to propagate annotations between pieces of an AV document. On the contrary we extend all of these relations into *conceptual* ones, considering every propagation (whatever its type) as contextual annotation.

Third, the possibility to consider AEs' attributs like representative images, scripts, digital features, *etc.* could narrow the distance between AE and *sem-cons* [11], or evolution of MPEG4 objects. We insist nevertheless on the *term* annotating the document rather than on the object on the screen. The term can of course be represented by an icon [10], but as a symbol, not as screen representation of real-world object.

Fourthly, the strata approach, hence the annotation interconnected strata approach, is not limited to audiovisual medium, but can be applied to any *sequential* medium, be it audio or text. This is the ability to freeze a part of audiovisual media (for instance an image), and to build relations between annotations that allow the description of non-sequential material.

Fifthly, the whole set of AVUs of a specific repository is tightly coupled to the corresponding knowledge base. So meta-models can express specific indexing or accessing methods depending on explicitly wanted strategies. On the other

relations this annotation has with other annotations (abstract ones in the knowledge base, or not). A shot can therefore be represented by an AVU u annotated with the AE $\langle Shot \rangle$: part of the meaning of u resides in this annotation, however other parts like structural shot-sequence relations, characters appearing in the shot, and others, are also in this annotation, in the sense that they are connected to it by some relation paths.

hand, generic user-oriented meta-models could be defined in order to help and orientate users in their tasks depending on the knowledge available on indexing strategies.

6 An AI-STRATA Based Tool for Indexing, Searching and Browsing Audiovisual Units

The aim of this part is to show how the AI-STRATA model is able to be an efficient base of conceptual modeling for different tasks and applications. As claimed in the introduction, any task exploiting an Annotated Audiovisual Document Base needs to some extend to describe audiovisual sequences. Our currently developped demonstrator includes the three main tasks to exploit an AADB:

The indexing task: this task involves the indexing of a sequence, where the person annotates with respect to some systematic procedure and anticipates later uses of the sequence. Systematic procedures are easy to describe (as meta-models) as predefined analysis dimensions, while anticipations of searches could be helped by "sounding" the AADB to retrieve how similar AVUs have been used (*i.e.* described for retrieval). We have also elaborated [19] the notion of *annotation assistants* as agents meant to help users to annotate sequences. The first and useful annotation assistant deals with automatic processing of video streams (cut and motion detection, similarity features extraction), leading to automatic ("visual" AE) or/and semi-automatic (eg. linking AE representating shapes/colors with AE denotating concepts) annotation of AV sequences. Other assistants could be added: sound annotation assistant, text-based annotation assistant, *etc.*

The search task: anyone searching for something is asked to describe what they are searching for. Ways to describe an AVU are infinite. We see the search process as an iterative process starting from a first description which depicts the first idea of the user, and is progressively enriched by results of the corresponding queries to the AADB. This process can be long and complex, but AI-STRATA should improve it in several ways by interconnecting first AEs with the available network of AEs (filters, contexts,*etc.*). Using local experience (*i.e.* previous searching episodes) or/and local examples (*i.e.* a local library of AVU) to elaborate the first description could greatly accelerate this process.

The browsing process: this is not exactly a task but a common way of searching documents by going from one point to another. Using AI-STRATA is straightforward for this activity. Starting from a first (weak) description the user can exploit existing AE relations from any retrieved AVU to slide from strata to strata. The ability to figure the starting typed links of each AVU as visual paths on a screen, and (in some way) to navigate at the knowledge level [15] gives more freedom to the person browsing and increases his/her interaction.

The editing process, that is the task of reusing and creating new audiovisual document taken from AV databases, can also rely on AI-STRATA model. Video abstracting, or storytelling can use annotations, and links between them to describe and find (as UAV) desired story units [8] and pieces of a document.

Editing could then just be reorganizing AVUs and annotations through a story model.

All these tasks can be long, complex and tedious. Exploiting experience to establish "winning" descriptions should be an efficient assistance. The Case Based Reasoning paradigm [1] will be used to help the description process by reusing past similar descriptions as bases for new description.

7 Conclusion and Future Work

We have presented in this article AI-STRATA, a new approach for audiovisual document description. Main contributions of the paper are: the proposition of several criteria for characterizing audiovisual representation approaches (time granularity, complexity, kind of characteristics, structure) ; the Annotation interconnected strata as an original way to describe AV documents taking into account both the dynamic dimension of AV streams thanks to atomic annotation elements, and the necessary structuring of the representation to cope with the inherent complexity of audiovisual material ; a generalized notion of context, based on the existing annotation, which should allow temporal, structural and conceptual contexts to be considered in the same way and which will be useful for browsing and annotation propagation. We consider also that every task related to AV retrieval systems is in fact a description task, and we illustrate how AI-STRATA is a promising way to easy out these tasks.

Current work to integrate the AI-STRATA approach has several objectives: develop an annotation application with the first annotation assistants (JAVA/ Corba based) ; develop mechanisms of contextual inferences leading to contextual annotation ; design and test the first meta-models with I.N.A. and France3 specialists ; and map the AI-STRATA model with data model proposed by database searches (semi-structured databases look promising). These research themes are the subject of "transversal" research in collaboration with other teams of the Sesame project.

All these developments have now become possible thanks to the unifying nature of the proposed model. We actually claim that most exploitation of audiovisual material can be expressed through the AI-STRATA model.

References

1. A. AAmodt and E. Plaza. Case based reasoning: Foundational issues, methodological variations, and system approaches. *AICOM*, 7(1):39–59, Mars 1994.
2. S. Adali, K.S. Candan, S. Chen, K. Erol, and V.S. Subrahmanian. Advanced video information system: Data structures and query processing. *Multimedia Systems*, 4:172–186, 1996.
3. P. Aigrain, D. Petkovic, and H.J. Zhang. Content-based representation and retrieval of visual media : A state-of-the-art review. *Multimedia Tools and Applications special issue on Representation and Retrieval of Visual Media*, 1996.
4. N. Belkin. Braque : Design of interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3):29–38, 1993.

5. S.F. Chang, J.R. Smith, M. Beigi, and A. Benitez. Visual information retrieval from large distributed online repositories. *Communications of the ACM*, 40(12):63–71, Dec. 1997.
6. T.-S. Chua and L.-Q. Ruan. A video retrieval and sequencing system. *ACM Transactions on Information Systems*, 13(4):372–407, October 1995.
7. J. M. Corridoni, A. Del Bimbo, D. Lucarella, and H. Wenxue. Multi-perspective navigation of movies. *Journal of Visual Languages and Computing*, 7:445–466, 1996.
8. G. Davenport. Indexes are out, models are in. *IEEE Multimedia*, pages 10–15, 1996.
9. G. Davenport, T. Aguierre Smith, and N. Pincever. Cinematic primitives for multimedia. *IEEE Computer Graphics and Applications*, pages 67–74, Jul. 1991.
10. M. Davis. Media streams: An iconic visual language for video annotation. In *Proceedings of the 1993 IEEE Symposium on Visual Languages*, pages 196–203, Bergen, Norway, August 1993. IEEE Computer Society Press.
11. W.I. Grosky. Managing multimedia information in database systems. *Communications of the ACM*, 40(12):73–80, Dec. 1997.
12. A. Gupta, S. Santini, and R. Jain. In search of information in visual media. *Communications of the ACM*, 40(12):35–42, Dec. 1997.
13. R. Jain. Visual information management. *Communications of the ACM*, 40(12):31–32, Dec. 1997.
14. R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Communications of the ACM*, 40(12):55–62, Dec. 1997.
15. J. Nanard and M. Nanard. Adding macroscopic semantics to anchors in knowledge-based hypertext. *Int. J. Human-Computer Studies*, 43:363–382, 1995.
16. C. Nastar, M. Mitschke, C. Meilhac, and N. Boujemaa. Surfimage: a flexible content-based image retrieval system. In *ACM Multimedia 98*, Bristol, Sept 1998.
17. E. Oomoto and K. Tanaka. Ovid: Design and implementation of a video-object database system. *IEEE Transactions on Knowledge and Data Engineering*, 5(4):629–643, Aug. 1993.
18. F. Pereira. Mpeg-7 : A standard for content-based audiovisual description. In *2nd Int. Conf. on Visual Information Systems*, pages 1–4, San Diego, Dec. 1997.
19. Y. Prié, J.-M. Jolion, and A. Mille. Sesame: audiovisual documents conceptual description model and content annotation assistants. In *CORESA'98, 4th. Workshop on COmpression and REpresentation of Audiovisual Signals*, Lannion — France, Juin 1998. CNET — France Télécom.
20. J.-F. Sowa. *Conceptual Structures — Information Processing in Mind and Machine*. Addison-Wesley, 1984.
21. R. Weiss, A. Duda, and D.K. Gifford. Composition and search with a video algebra. *IEEE Multimedia*, 2(1):12–25, 1995.
22. B.L. Yeo and M.M. Yeung. Retrieving and visualizing video. *Communications of the ACM*, 40(12):43–52, Dec. 1997.
23. K. Zettsu, K. Uehara, K. Tanaka, and N. Kimura. A time-stamped authoring graph for video databases. In *Databases and Expert Systems Applications, LNCS 1308*, pages 192–201. Springer-Verlag, 1997.