

Generating Adaptive Multimedia Presentations Based on a Semiotic Framework

Osama El Demerdash, Sabine Bergler, Leila Kosseim¹ and PK Langshaw²

¹ Concordia University, Department of Computer Science and Software Engineering

² Concordia University, Department of Computational Arts

Abstract. We propose a framework for generating adaptive multimedia presentations through the dynamic selection of files from a large data repository. The presentation is generated based on the technical (syntactic), semantic and relational textual annotation of the data as well as context-sensitive rules and patterns of selection discovered with the aid of the system during the preparation phase. We borrow concepts from the fields of discourse analysis and rhetorical structure as the theoretical basis of our work. To validate the framework, a prototype was developed using Java, Flash-MX and XML.

1 Introduction

A performer, with a considerable repository of multimedia material to support her presentation, consisting of approximately 2,000 files divided between images, video, animations, voice, sound and music excerpts may wish to enhance her performance by relying on a system to dynamically generate presentations through selecting and playing the most appropriate material. The system should do this based on the context of the performance and upon a trigger from the performer. The conceptual presentation is an abstraction of what the performer has in mind as a general idea of her presentation. During the actual presentation, the performer might intentionally decide to deviate from the original plan, by visiting related themes or raising new arguments, or may find herself drawn into new areas as a result of the interaction with the audience. The role of the system is to adjust to the actual presentation context and provide just-in-time support material during a performance or in the preparation/rehearsal phase, allowing different alternatives to be assessed. The system can also interact directly with a spectator, to produce a personalized presentation.

1.1 A Semiotic Perspective of Multimedia

Multimedia is becoming increasingly accessible and diffusible on the WWW. More and more applications are being developed to process multimedia objects, generally requiring storage, indexing, retrieval and presentation of multimedia. Much research in this area deals with content-based retrieval, the automatic recognition of the content of the medium [1]. However, modeling of the data and task is often biased toward information retrieval, incorporating temporal and spatial models, but ignoring other contextual and relational factors.

Systemic Functional Linguistics (SF) O’Toole demonstrates through the analysis of a painting [2] that Systemic Functional linguistics [3] is broad enough to cover other semiotic systems, particularly visual ones. In his analysis, the different constituent functions of the model (ideational, interpersonal and textual) are projected over the representational, modal and compositional functions in the visual domain. In the Systemic Functional model, text is both a product and a process. Language construes context, which in turn produces language [3]. In the light of this theory, it is possible through analysis to go from text to context, or through reasoning about the context to arrive at the text — though not the exact words — through the triggering of the different linguistic functions. While we do not try to draw exact parallels between the Systemic Functional model as applied in linguistics and in multimedia, we retain some of the highlights of this theory; most notably the relation between text — in our case multimedia — and its context.

Rhetorical Structure Theory (RST) We also draw on Rhetorical Structure Theory (RST) [4] for representing the possible relations between the different components of the model. RST has been used to analyze the relations between text spans in discourse and to generate coherent discourse. RST analyzes different rhetorical and semantic relations (ex. precondition, sequence, result) that hold between its basic units (usually propositions). In our framework, we used RST as a design solution to guide us in the planning of a coherent performance modeling the relations among multimedia data similarly to the relations among text spans in a discourse.

We adapted this theory to multimedia and artistic applications by defining new relations which reflect the implicit artistic processes.³ This included a representation of more than one level of interpretation to account for the sometimes intentional ambiguity of art; contrary to technical discourse, the artistic expression creates a more open environment encouraging different interpretive possibilities.

2 Related Work

Multimedia applications are task, domain, process or media dependent. Due to the resulting complexity it is necessary for any framework/model to strike a balance between generality and applicability. Jaimes [5] describes a visual information annotation framework. The MATN (Multimedia Augmented Transition Network) by Chen et al. [6] proposes a general model for live interactive RTSP (Real-Time Streaming Protocol) presentations, which models the semantics of interaction and presentation processes such as Rewind, Play, Pause, temporal relations and synchronization control (e.g. concurrent, optional, alternative), rather than the semantics of the content. In the HIPS project modeling the context,

³ Currently, the prototype implementation does not model these relations, they were rather used manually during the conceptualization phase.

user and their interaction for museum's guides [7] and [8], a portable electronic museum guide transforms audio data into flexible coherent descriptions of artworks that could vary with the context. The system uses the *MacroNode* approach, which aims to develop a formalism for dynamically constructing audio presentations starting from atomic pieces of voice data (macronodes) typically one paragraph in length. A museum visitor, could get one of several realizations of the description of an artwork depending on the context of interaction. The context is defined according to the visitor's physical location in relation to the described artwork. In this approach, the data is annotated with the description of content and relations to other nodes. These relations are conceptually similar to relations in Rhetorical Structure Theory. In a later project by Zancanaro [9], the utilization of RST relations is extended to producing video like effects from still images, driven by the audio documentary. A closed-set ontology is employed. Kennedy et al. [10] developed a communicative act planner using techniques from Rhetorical Structure Theory (RST). The purpose of the system is to help animators by applying techniques to communicate information, emotions and intentions to the viewer. The knowledge base of the system includes information about scenes, shots, space, time, solid objects, light, color, cameras and cinematographic effects. The tool is intended to be used in the planning phase to alter a predefined animation in a way perceptible to the viewer.

These systems were not designed to be fully adaptive to changing context and are domain or task specific.

3 An Adaptive Framework

Figure 1 is an illustration of the components of our adaptive multimedia framework. *Static Components* refers to elements not contributing directly to the adaptive potential of the framework, i.e. fixed for different presentations. *Dynamic Components* are responsible for the adaptive aspect of the framework, which is designed to be general enough for different contexts, complex presentation requirements requiring more elaborate utilization of the framework. A partial proof of concept implementation of this framework will be described in Section 4, including the data model, the information retrieval model and limited areas of the context model, selection heuristics and effects. Here, we present all envisioned alternatives as design solutions.

3.1 The Data Model

Our data model consists of the data files (in any format supported by the visualization software⁴) and their annotations with technical and semantic features as well as relational characteristics. Following Prabhakaran [11], we model multimedia objects as a general class with specialized classes for each type of media.

Meta-data describe semantic features of the media files and are constant across media types. These include *Keywords* describing the semantic content

⁴ We currently use Flash-MX which supports formats including mp3, mpg, swf, html...

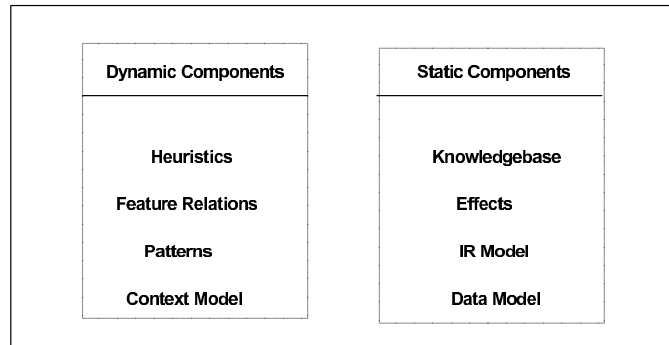


Fig. 1. Framework for Adaptive Multimedia Presentations

and the Mood of the selection. Defining these subjective annotations is part of the creative process of performance design and may be more or less useful for another performance. To achieve reusability, general classification ontologies can be used such as distinguishing between *General*, *Abstract* and *Specific*. *General* refers to a class of objects with physical presence like *human*, *chair*, *dog*. *Abstract* is an idea or concept without a physical presence like *hunger*, *war*, *sleep*, and *Specific* is a subclass of *General* for identifiable named entities.

This model is extensible through the use of any relevant ontology, since the annotations are cumulative. For instance, in the current project we include a feature called *Mental Space* with the attributes *dream/reality/metaphoric* and another feature *Physicality* to convey relative size of objects with the attributes *landscape/body/page*. These features may not be relevant to all performances, but may be interesting to some.

Each type of media is also annotated according to its specific characteristics as illustrated in the remainder of this section.

Text To delimit the medium ‘text’ is a tricky task. Audio data may contain spoken text, images may contain text fragments, complete poems could be laid out in a visual way (as frequently done by Apollinaire), and video and animation may include both spoken and visual text. We define *text* as textual data formatted in ASCII format (e.g. txt, rtf, HTML). Text is the most researched medium in the field of information retrieval and we employ tools similar to many modern search engines, based on prior indexing of keywords.

Image Still images are digital graphics containing drawings, paintings, photographic images, text or any combination of these. Technical features commonly used for annotating graphics include color, texture, dimensions and file format. In this project texture was deemed irrelevant and excluded. Multiple color annotations were permitted. Sequences of still images are handled using relations

between specific files, or through retrieving by patterns, with control over certain features (e.g. speed) through the user interface.

Moving Images Moving images include videos and animations. We use atomic excerpts consisting generally of a few seconds to two minutes. This roughly corresponds to the definition by [12] of a *Scene*. A *Scene* is a collection of contiguous logically related shots, while shots are contiguous frames with common content. Sequences, which form a higher level in this hierarchy, are not considered as units, but are dealt with through relations. This category has a time dimension, represented by the *duration* and the *pace* features. The choice of *Scene* as the basic unit enables the generalization of image features to the video excerpt. For example, *Color* is the dominant color in the scene.

Audio Audio data is classified as *music*, *speech*, or *other* sound data (e.g. Electro Acoustic, noise etc.). In addition, temporal features like *duration* have to be indicated. For music, we also indicate *pace* (tempo) using qualitative attributes *fast/medium/slow*, and *type* (*melodic*, *harmonic*, *percussive*, *gestural*). Like text data, speech carries information in natural language.

Relations Relations between the media files are also annotated. As mentioned earlier, we use modified RST-like relations for two reasons. Firstly, to impose temporal constraints on the order of playing these files, in order to insure the production of a coherent and cohesive presentation. Coherence is achieved through the logical temporal and spatial ordering of the different selections of the presentation, while cohesion results from the synchronization of two or more selections. Secondly, to construct a relational navigation map linking the selections according to their sensory and/or semantic links. Multiple relations can be represented, forming a web rather than tree structure, which is customary - though not a requirement - even in the case of text structure [13]. Since RST relations are rhetorical in nature, we augment them with relations for *temporal constraints* (*follow*, *precede*, *simultaneous*) and others that express pure sensory associations (*phonetic*, *visual*).

Figure 2 shows the representation of the relations between media files in RST format.

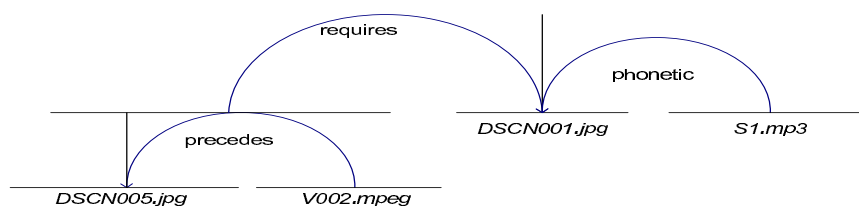


Fig. 2. RST-like relations representation

3.2 The Context Model

For multimedia presentations, we define context to include several interdependent features: the *outline*, *time*, *space*, *presenter*, *audience*, *medium*, *rhetorical mode*, *mood*, and *history*. This certainly does not represent a closed set; a more refined framework could of course use several other such features. We give here a brief description of the context variables.

The Outline The outline of a presentation corresponds to the Field of Discourse in SF. It is expressed in terms of keywords. Like the outline of an essay, or a book's table of contents, a presentation outline is a representation of its plan. The user is able to change the subject through the interface of the system, triggering a system response in the form of new material related to the current subject. Outlines could be complex and contain overlapping sections. The outline can be ordered or unordered and must support time constraints as needed, as well as a weighting scheme to indicate the relevance, or relative importance of each keyword in a section. Rhetorical modes and moods for each section should be supplied also.

Time and Space The intended timeline is a determining factor in the planning of the presentation, since it is used to avoid overflows and empty gaps, as well as to balance media selection. Overflow can happen when a certain media selection, for example a video relating to a particular topic in the outline, turns out longer than that section's initially planned time-slot. Conversely, empty gaps occur when there is not enough material to fill the allotted time for a particular topic. Balancing media selection can help generate more appealing presentations and requires keeping track of time. Time can be modeled at the required level of accuracy (min., sec., etc...). The capability to relate Time and the Outline of the presentation and to dynamically change this relation is important.

The physical size of the performance space as well as its placement (inside/outside) provide hints to the appropriate type of media to play. Presets can handle different space configurations.

We define *virtual space* to be the spatial layout on the screen, relevant when multiple objects are presented on the screen simultaneously or when one object does not occupy the full viewing space.

The User Profile Despite their correlations, the user model is often considered separate from the context model in application design. We chose to include the user profile in the context model. Indeed, the presenter and the audience together correspond to tenor in SF. A user profile can be represented as keywords, preferably drawn from the different ontologies applied in the data model to avoid an extra step of matching terms. Other user profiling techniques include registering the users' requests to determine their interests.

Audience Gender, age, background and relationship to the author of the presentation are all potential selection factors. For example, children might be more responsive to images and animations than to text and video. Artistic, scientific and multidisciplinary audiences require different communicative strategies, for instance for presenting from a position of authority as opposed to a peer-to-peer presentation. Employing stereotypes has become a common practice for modeling anonymous audiences, especially in web-based applications which service a significant number of users with varying characteristics and interests.

Media In the context model, we consider video, audio, animation, image, text and combinations of these, whether simultaneous or overlaid. These media types should not be confused with the medium attribute in the data model, which is used to characterize the medium of single files, or that in the query specification which can be used to constrain the types of media in the result set. The currently playing media types are an essential context parameter and are used by heuristics such as to decide whether or not to interrupt the current selection.

Rhetorical Mode The rhetorical mode is the communication strategy used at a given moment in the presentation to affect the audience in particular ways. Examples of rhetorical modes given by Halliday include *persuasive*, *expository* and *didactic* [3]. There is no consensus on rhetorical modes in the literature on essay writing, however more modes are usually considered including among others Narrative, Descriptive, Illustrative, Comparison/Contrast, Process analysis, Definition and Cause/Effect.

Moods The emotional feel of the presentation or its mood contributes to maintaining a coherent context. Moods could either be directly mapped to elements in the taxonomy of the project, or explicit links could be established through the use of feature relations and heuristics. The definition of Moods themselves has to be qualitative and may be comparative (e.g. *happier*, *happy*, *neutral*, *sad*, *sadder*). Color psychology establishes relationships between colors and moods. For example, Red is often associated with anger and excitement, blue with sadness and calm, green with nature, envy etc. However, other properties of color such as hue and saturation also affect the mood. In music, loudness, rhythm and key are all factors affecting the mood.

History A record of selections already played should be used to avoid repetition of these selections. History could also be used to balance, as desired, the concentration of the different media in the presentation and to diversify the selection as required. Moreover, it is possible to use the history to reproduce a presentation, or as training samples for machine learning techniques.

3.3 Feature Relations

Relations are used either at the level of individual data files to link selections together as described in the previous section, or at the abstract level. When used as such, they serve to establish explicit relations between the different features of the data model, providing for overriding capabilities, and thus an additional interpretive layer. These relations could be applied within the same medium, for example associating a certain color with a mood, or across different media types, such as yellow with jazz music. Feature Relations can also be used to express constraints, which can be considered as negative relations. For example, to express that Loud music should not accompany Calm mood.

3.4 Heuristics and Experiments

The goal of the selection heuristics is to produce different interpretations of the performance, according to the context, and through the selection and ordering of multimedia material. The process involved is a context-to-content mapping. The context of the performance, in addition to any explicit triggers, is mapped into specific selections. To refine the relevant heuristics, experiments should be conducted during rehearsals through variations of the different features.

During the presentation, the system will keep track of the current section, of selections played, of changes in communicative goals (e.g. the performer decides to give more time to presenting a certain section or adds a new section) and ideally will offer the user to trigger, browse or query the media using a visualization tool convenient for the criteria specified above. When the performer digresses or changes one of the presentation context parameters, the system should be able to play, based on the given heuristics for the given context, an appropriate selection. An example might be giving weights as follows: $\text{Remaining_time} \times .2 + \text{Mood} \times .3 + \text{key_words} \times .3 + \text{Main_goal} \times .2$. The system will then use this equation to search for the most appropriate selection.

3.5 Visual Effects

Visual effects are techniques used in the presentation model to improve the visual quality of the presentation. They are also used to enhance the relation between two selections in the presentation for example by associating a certain kind of relation with a transition. Effects are applied to alter images, and do not create new ones. They include transitions (*cut, fade-in, fade-out, dissolve, wipe*), scaling, zooming, layering etc. These effects are commonly available in the design-mode of presentation software like MS-PowerPoint and Macromedia-Flash, or through programming. However, including them in the run-time interface in an accessible manner, allows the presenter to apply them on the fly during the presentation. The application of visual effects has a long tradition in fields such as cinematography where transitions roughly correspond to punctuation in language.

3.6 Generation Patterns

Patterns are recurring designs, behavior and conditions. In the context of our framework, Patterns could be formed of complex combinations of features and heuristics. Generation patterns are discovered while experimenting with the system during the rehearsal/preparation phase of the presentation. Once identified and included in the interface, they can be retrieved explicitly during the presentation. For example, a Surprise pattern could be a combination of loud dynamics, fast video, and a set of heuristics that changes fast across the different media and colors. This complex goal would be difficult to achieve otherwise in real-time. Defining patterns can also lead to more meaningful ways of describing the higher level goals of the presenter.

3.7 The Information Retrieval Model

A framework for adaptive multimedia presentations must include an information retrieval component, since pre-arranging all possible combinations of media would be infeasible in large repositories. The information retrieval model defines the way the selection criteria are applied to the annotated data to determine the relevance of documents. The most popular model in use nowadays for text retrieval is the Vector model. Other models include the Boolean and Probabilistic models. For a thorough discussion of information retrieval and the different models see [14].

3.8 The Knowledge Base

While some expertise already exists in each medium separately, there is no evidence of standardized practices in the creation of an adaptive multimedia presentation. Once the expertise in the domain of multimedia presentations has been developed, it is beneficial to capture this expertise and exploit it in a systematic manner. The Knowledge base would act as a permanent repository of this expertise. Such expertise might include for example techniques, feature relations, heuristics, ontological hierarchies of strategies, meanings, effects and rhetorical relations.

4 Implementation

A prototype of the proposed framework has been constructed. It was developed using a three-tier software architecture on flash/java/MYSQL platforms illustrated in Figure 3. The model tier consists of the data and annotations, relations and retrieval patterns. Business logic including operations on the database, heuristics and status information makes for the middle tier, while the presentation (view) tier has the user interface and interaction elements. Long-term experimental goals, the volatile nature of requirements and other practical considerations have influenced the three-layered, modularized architecture. Separating

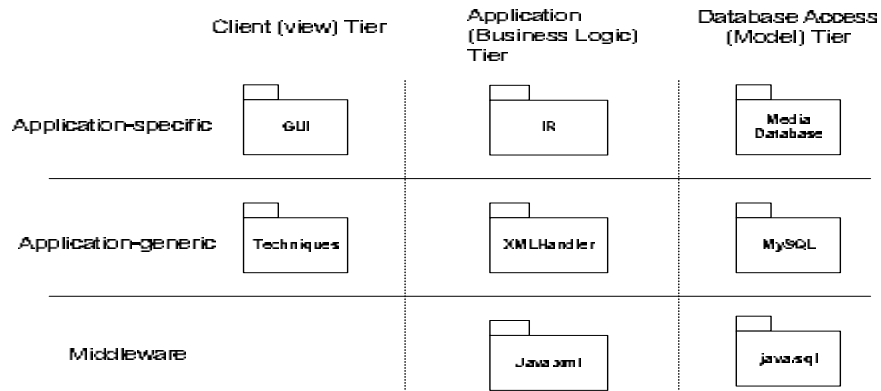


Fig. 3. Architecture of the System

the presentation from the model and the business logic permits the substitution of any of these layers at minimum cost.

Figure 4 illustrates the interaction sequence for retrieving data , with the following scenario: The user indicates through the presentation graphical user interface (GUI) the features of the data to be retrieved. The presentation GUI reproduces the user's request in XML format and sends a message to the XML handler to process the request. XML sent by the presentation GUI to the XML handler includes elements for both <sound> (audio) requests and <content> (visual) requests. The XML handler forwards the request to the Query Processor. The Query Processor applies heuristics relevant to the required features. The Query Processor constructs a SQL statement according to the requested features and heuristics and runs it on the media database. The media database returns the result set to the Query Processor. The Query processor translates the result set into XML format and sends it to the XML Handler. The XML Handler forwards the XML result set to the Presentation GUI. The <Slides> element represents visual files while the <Sounds> element represents audio files. It is necessary since the Presentation GUI deals with these categories separately and

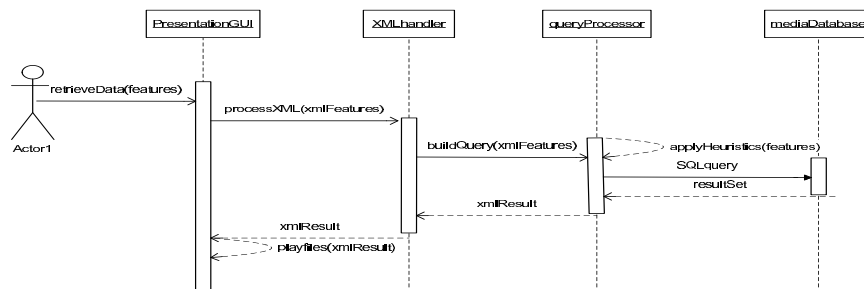


Fig. 4. Sequence Diagram for Retrieving Data

in different manner. The presentation GUI displays the files specified in the result set.

The interface is used for informing the system of changes in the presentation/audience models and to control/override the system's suggestions using relevance feedback and possibly navigation of the knowledge base.

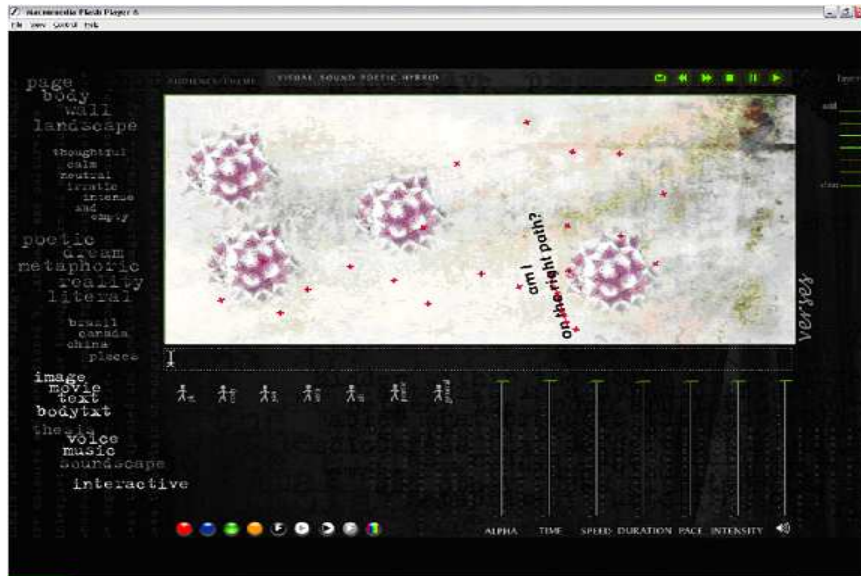


Fig. 5. Screen-shot of the system's interface

Figure 5 shows a screen shot of the system. The user can select any of the direct features (*time*, *spectrum*, *alpha*) through the presentation GUI. They are either linked internally to the context and data models or generate visual effects. Using the relations (Section 3.3) and the user specifications, the most appropriate data files are retrieved from the multimedia database. Of the relevant data files, only a subset may be used in the final presentation. The final selection and ordering for the presentation is made by the selection heuristics and the generation patterns which ensure a coherent final presentation.

5 Evaluation

The presented framework has to be evaluated over time on mostly qualitative reusability measures, such as applicability, scalability and ease of adoption. Our prototype implementation using this framework is a first attempt. Here, we are mainly concerned with identifying and implementing benchmarks to evaluate similar systems. Common methods for evaluating information retrieval systems

focus on measuring the effectiveness of the system [15], defined as the relevance of retrieved documents as determined by a human judge. Precision and Recall figures are then calculated for the system according to the following formulae:

$$Precision = \frac{\# \text{ of relevant documents found by the system}}{\text{total \# of documents retrieved}} \text{ and}$$

$$Recall = \frac{\# \text{ of relevant documents retrieved by the system}}{\text{total \# of relevant documents in the collection}}.$$

Measuring Precision at different recall levels might give a better idea about the effectiveness of the system.

The influential Text Retrieval Conference (TREC) adopts this benchmark. However, [16] questions this measure, pointing out that Precision and Recall measures ignore such important factors as the relativity of the relevance of a document. The binary division into relevant and non-relevant documents is an oversimplification. [14] notes, Precision and Recall presume that an objective judgment of relevance is possible, an assumption readily challenged by the high interannotator discrepancies.

The case of Multimedia Information Retrieval offers other particularities and difficulties which need to be considered for evaluation. The TREC 2001 Proceedings, which included a video track for the first time, acknowledge the need for a different evaluation system for that track [17]. [16] lists some of the performance criteria not captured by Precision and Recall are speed of response, query formulation abilities and limitations and the quality of the result. He presents the notion of Approximate Retrieval (AR) arguing that unlike text data, the characteristic ambiguity of both multimedia information and queries could only lead to an approximate result, and asserts the significance of rank, order, spread and displacement in Multimedia Information Retrieval. Schauble [15] introduces a notion of subjective relevance which hinges on the user and her information needs rather than on the query formulation. Thus we feel that the user's participation in determining the relevance of the result is an essential factor.

Certain characteristics of our system add to the complexity of the evaluation task. These include the user model, context-sensitive retrieval, and the layer of subjective relations between features and/or elements in the data model introduced explicitly by the user.

We propose here to use alternative user oriented measures for the evaluation of the system. The first two of these measures, namely the Coverage and Novelty Ratios reported by [14] measure the effectiveness of the system with respect to the user's expectations. The Coverage Ratio measures the ratio of documents which the user was expecting to be retrieved over what was actually retrieved by the system:

$$Coverage = \frac{\# \text{ of relevant documents known to the user and retrieved}}{\# \text{ of relevant documents known to the user}},$$

while the Novelty Ratio is the ratio of relevant documents which were not expected by the user:

$$Novelty = \frac{\# \text{ of relevant documents retrieved previously unknown to the user}}{\text{total \# of relevant documents}}.$$

In order to apply these measures, a special evaluation environment needs to be set up to run the system in interrupted mode so that the user can evaluate the selections played without interfering with the system heuristics.

Finally, the audience could help evaluate the system from a different perspective: its higher-level goals of providing the audience an entertaining and informative experience. Interviews or questionnaires could be used for surveying the audience's reaction to the performance.

6 Conclusion and Future Work

Multimedia presentations provide a powerful tool for communication. However, in order to exploit the full potential of multimedia presentations, it is essential to allow for a certain flexibility in the selection of the material for a more dynamic presentation. We proposed a framework for adaptive multimedia presentations. The framework included models for data, context and retrieval, selection heuristics, retrieval patterns and multimedia techniques. As a proof of concept, we implemented a system that dynamically selects and plays the most appropriate selection of multimedia files according to the preferences and constraints indicated by the user within the framework. We borrowed concepts from text analysis to model the semantic dimension of the presentation. We also proposed adopting different methods of evaluation to reflect the subjective nature of the task. A prototype implementation has been successfully used for several performances and is the preliminary validation of our claims.

We see a need for developing a general architecture for multimedia run-time manipulation. Such an architecture should facilitate replacing and augmenting different components of the framework. Such an architecture must also include tools for accomplishing common tasks such as data annotation and interface design. The goal is a robust architecture, with acceptable scaling and generalization capacity. Some areas where we see possibilities for improvements and innovation are:

- Using other triggering mechanisms as speech, gesture or cluster-based content navigation maps to navigate through the concept space and visualize the relations between the concepts
- Developing an annotation tool to help the inexperienced user
- Investigating alternative architectures like agent-based architectures
- A web-based multi-user version would allow for collaborative production of performances in real-time.
- Using supervised machine-learning techniques, using relevance feedback for discovering heuristics and building user profiles. History data collected from performances can be used as training data. Relevance feedback can also be used in real-time during the performance as a corrective measure.

References

1. Flickner, M., Sawhney, H., Nublack, W.: Query by Image and Video Content: The QBIC System. In: Intelligent Multimedia Information Retrieval. California:AAAI Press/ The MIT Press (1997)

2. O'Toole, M.: A Systemic-Functional Semiotics of Art. In: *Discourse in Society: Systemic Functional Perspectives*. Ablex Publishing Corporation, New Jersey (1995)
3. Halliday, M., Hasan, R.: *Context and Text: Aspects of Language in a Social Semiotic Perspective*. Oxford University Press, Oxford (1989)
4. Mann, W., Matthiessen, C., Thompson, S.: Rhetorical Structure Theory and Text Analysis. In: *Discourse Description: Diverse Linguistic Analyses of a Fund-raising text*. John Benjamins Publishing Company, Amsterdam (1992)
5. Jaimes, A., Shih-Fu: A conceptual framework for indexing visual information at multiple levels. In: *SPIE Internet Imaging 2000*. Volume 3964. (2000) 2–15
6. Chen, S.C., Li, S.T., Shyu, M.L., Zhan, C., Zhang, C.: A multimedia semantic model for RTSP-based multimedia presentation systems. In: *Proceedings of the IEEE Fourth International Symposium on Multimedia Software Engineering (MSE2002)*, Newport Beach, California, ACM (2002) 124–131
7. Not, E., Zancanaro, M.: The MacroNode approach: Mediating between adaptive and dynamic hypermedia. In: *Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*. (2000)
8. Marti, P., Rizzo, A., Petroni, L., Diligenti, G.T.M.: Adapting the museum: a non-intrusive user modeling approach. In Kay, J., ed.: *User Modeling: Proceedings of the Seventh International Conference, UM99*, Banff, Canada, Springer Wien New York (1999) 311–313
9. Zancanaro, M., Stock, O., Alfaro, I.: Using cinematic techniques in a multimedia museum guide. In: *Proceedings of Museums and the Web 2003*, Charlotte, North Carolina, Archives and Museum Informatics (2003)
10. Kennedy, K., Mercer, R.: Using communicative acts to plan the cinematographic structure of animations. In Cohen, R., et al., eds.: *Advances in Artificial Intelligence: Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI2002*, Calgary, May 27-29, Springer, Berlin (2002) 133–146
11. Prabhakaran, B.: *Multimedia Database Management Systems*. Kluwer Academic Publishers (1997)
12. Carrer, M., Ligresti, L., Ahanger, G., Little, T.D.: An Annotation Engine for Supporting Video Database Population. In: *Multimedia Technologies and Applications for the 21st Century: Visions of World Experts*. Kluwer Academic Publishers (1998) 161–184
13. Marcu, D.: *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA (2000)
14. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. ACM Press and Addison Wesley (1999)
15. Schäuble, P.: *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*. Kluwer Academic Publishers (1997)
16. Narasimhalu, A.D., Kankanhalli, M.S., Wu, J.: Benchmarking Multimedia Databases. In: *Multimedia Technologies and Applications for the 21st Century: Visions of World Experts*. Kluwer Academic Publishers (1998) 127–148
17. Smeaton, A.: The TREC 2001 video track report. In Voorhees, E., Harman, D., eds.: *The Tenth Text Retrieval Conference, TREC 2001*. NIST Special Publication 500-250, NIST, Gaithersburg, Maryland (2001) 52–60