INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

A CONCEPTUAL FRAMEWORK FOR ADAPTIVE MULTIMEDIA PRESENTATIONS

Osama El Demerdash

A thesis

in

The Department

of

Computer Science

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Computer Science

Concordia University

Montréal, Québec, Canada

September 2004

0-612-94738-6

Abstract

A Conceptual Framework for Adaptive Multimedia Presentations

Osama El Demerdash

In this thesis we propose a framework for a system that dynamically selects and plays multimedia files from a large data repository in order to produce a presentation. The presentation is generated based on the technical, 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 with data created and annotated by a research group from the Department of Design Art.

Acknowledgments

I would like to express my deepest gratitude to my supervisors Drs. Leila Kosseim and Sabine Bergler for their guidance, support, and patience, as well as Prof. PK Langshaw from the Design Art department for her support, inspiration, and feedback. I would also like to thank all the artists in Prof. Langshaw's group who have contributed invaluably to my understanding of the artistic processes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Even the Catholic Church of the middle ages was tolerant by modern standards.
George Orwell, 1984

1.1 Problem Definition

A performer, with a considerable repository of multimedia material to support her presentation, may wish to enhance her performance by relying on a system to dynamically select and play/generate the most appropriate material. The system should do this based on the context of the performance and upon a trigger from the performer. Examples of the context of such presentations could be spoken artistic performances, classroom presentations or dynamic museum guides.

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 and the questions they might raise. The role of the

system is to keep track of the actual presentation context and be able to provide just-in-time supporting material, using its knowledge base and by applying appropriate rules concerning its involvement in the presentation. The system also needs to keep track of the history of its interaction with the user.

The system needed to tackle the problem can be viewed as an intelligent multimedia information retrieval system, which according to Maybury et al. "lies at the intersection of artificial intelligence, information retrieval, human-computer interaction, and multimedia computing." [Maybury, 1997] There is space for research in each of these individual areas as well as in their integration in a single system.

The system might also be thought of as an assistant to the director of the performance, which could change the selections played according to the responsiveness and the profile of the audience as well as that of the performer. It could also act independently to produce a linear presentation or interact directly with the spectator as a user in the absence of the performer, allowing her to produce a personalized presentation. Finally the system could have a role in the preparation/rehearsal phases of the performance, when different alternatives will be assessed to help reach an optimum model of the presentation.

1.2 Research Motivation

Traditionally, information science adopts a rather structuralist approach to arts, seemingly to avoid complexity and to account for technical obstacles. This outlook coincides with the typical motivation for using computers: efficient mass production. While this motivation is valid for commercial applications, it proves rather alienating in the artistic ones. We can group most applications in the art fields into two distinct

groups: ones which are "scientifically oriented" and the others which are "artistically oriented".

Scientifically oriented successful attempts in the arts domain focus frequently on the more tangible applications and phases of the artistic experience. Examples include automated museum guides [Not and Zancanaro, 2000] and virtual models of art and culture [Champion, 2003] and [Chan, 2003]. Such applications, often based on well-founded scientific methodologies and representing models of the real world, are susceptible of proof, validation and evaluation with scientific rigor. In their quest to objectivity, they also strive to present general reusable solutions to well-defined problems. This scheme subscribes to efficiency requirements, omni-present in the field of information technology. Such work can be characterized as conclusion oriented. It places more emphasis on reaching a conclusive quantifiable outcome than on dealing with the nuances of the artistic content.

On the other hand, artistically oriented applications cater to the production of tools such as editing tools and tools for generative and hybrid art; art which makes use of sensors and other electronic devices and investigates virtual reality environments. As such, it can be considered technology adapted to art. Examples of such applications include animation and digital imaging and sound tools.

With this project, we aim to integrate Art and Technology so that neither is corollary to the other. Adaptive multimedia presentations involve both a preparation and a production phase. We use technology throughout the production from its conception through rehearsal to the final production. Our main objective is to produce a framework for dynamic presentation which can be significantly and tangibly different

from the spectators' point of view; reproduced from a fixed set of multimedia components including text, images, music, animations and videos. Moreover, we seek to avoid relying on random criteria for selection, instead using the context of the performance and the performer's preferences to empower her with a set of rich, predictable tools for manipulating the performance.

1.3 A Semiotic Perspective of Multimedia

As multimedia is becoming increasingly accessible and diffusible on the WWW to the average user, more and more applications are being developed to process multimedia objects. This processing generally consists of storage, indexing, retrieval and presentation of multimedia. Much of the research in this area deals with the technically thorny and yet-to-be-resolved task of content-based retrieval. Content-based refers to the automatic recognition of the content of the medium. (ex. [Smeaton, 2002]). However, the modeling of the data and the task is often secondary or handled in a similar fashion to textual data, without regard to the rich and complex nature of the information conveyed by diverse media. While temporal and spatial models are sometimes incorporated, other contextual and relational factors are ignored.

1.3.1 Systemic Functional Linguistics (SF)

Indeed, just considering that we use more senses for interpreting multimedia data calls for a different approach for modeling multimedia retrieval and presentation tasks. Since the production we are dealing with is anchored on text, we have chosen to build on theories and representations from linguistics and computational linguistics, which

could also be applicable, after certain modifications, to other media. We found inspiration in Systemic Functional Linguistics as described in [Halliday and Hasan, 1989]. As O'Toole illustrates through the analysis of a painting [O'Toole, 1995], the Systemic Functional model 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 [Halliday and Hasan, 1989]. 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.

1.3.2 Rhetorical Structure Theory (RST)

We also draw on another linguistics theory widely used nowadays, namely Rhetorical Structure Theory (RST) [Mann *et al.*, 1992] for representing the possible relations between the different components of the model. RST has been used as a tool to analyze the relations between text spans of a discourse; but also as a tool to produce a coherent discourse. RST endeavors to analyze texts based on the different rhetorical and semantic relations within its basic units; usually at the propositional level. By selecting text spans that hold certain semantic and rhetorical relations among

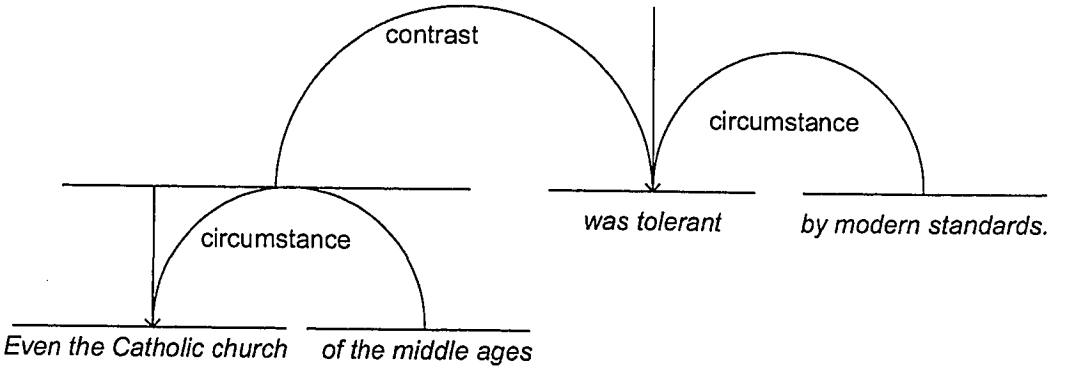

Figure 1: RST in text analysis

themselves (ex. precondition, sequence, result), it is possible to generate a coherent discourse from various text components. The span containing the main claim in the relationship is referred to as *nucleus*, while the evidence of the claim is described as *satellite*. Due to complexity and ambiguity, it is sometimes possible to arrive at two different analyses for the same text, even by the same annotator. In our framework, we used RST as a design solution to guide us in the production of a coherent performance; as the relations among multimedia data is seen similarly to the relations among text spans in a discourse. Figure 1 shows an example of using RST relations in text.

However, in order to adapt this theory to multimedia use, as well as to artistic applications, we identified the need for some changes. (At the current phase, the prototype implementation does not incorporate these relations, but they are included in the discussion.) For example, the implementation of new relations which reflect the artistic processes of *inspiration* and *association*, implicit and inherent in the arts domain. Also, the recognition of the need to represent more than one level of interpretation to account for the sometimes intentional ambiguity of art; contrary to technical discourse, the artistic language provides for a more open environment

encouraging different interpretive possibilities.

Such aspects are often overlooked as the goal to appeal to scientific standards and culture prevails, resulting in comparative and practical systems. Our aim is to strike a balance between the scientific tradition of objectivity and the highly subjective nature of media art. In order to do this, we have to take into consideration, in addition to the content and structural relations of the performance, its contextual variables such as the planning phase, space, the performers mental model and the audience, all of which might be at least as relevant as the content of the performance.

We also use technology during the rehearsal phase in an attempt to identify meaningful artistic patterns, which can be recalled easily during the performance. Although we are mostly considering trial and error strategy during the current phase of the project, we could apply machine-learning methods in the next phases to learn the system's parameters. This, in addition to facilitating the performer's task, might shed light on relations between the different media and forms, and help in cognitive research of the artistic process.

Finally, we turn to philosophy for a natural source of foundation and validation of our premises. Of special interest to us is the philosophy of interpretation (Hermeneutics). Indeed, as we set out to work on the project we felt the need for an interpreter to reveal the hidden, ungraspable and ambiguous differences in meaning between scientific and artistic discourse. To a certain extent, philosophy can be considered the least common denominator of Arts and Science.

1.4 Premises

This project is the result of a collaboration effort between a research group from the department of Design Art, led by PK Langshaw and CLaC, the NLP group led by Bergler and Kosseim. A prototype of the implementation was presented at VSMM 2003 [Demerdash *et al.*, 2003]. The "artistic" group comprised visual artists, musicians and wordsmiths. Material was produced independently by members and subgroups of the group then passed along to other group members for response. The produced media database consists of approximately 2,000 files, divided between images, video, animations, voice, sound and music excerpts, with images making up the bulk of the material. Figure 2 shows the cycle of interpretation involving the artists, the performer/narrative designer and the audience. In this figure, the cyclic arrow from the artist group refers to the artistic creation process, where the artists took one another's work and interpreted it in a different medium. Table 1 illustrates the scope and representation of these interpretation in the suggested framework. In this case, the scope is the artist group and the interpretations are represented in the framework as relations equivalent to rhetorical relations in text as described earlier (see section 3.1.6 for example of relations).

The second interpretation in chronological order of the performance involves the performer interpreting the production of the artist group. In fact, the performer does a pre-interpretation of the artistic work during the annotation phase. This is represented in the framework by the dynamic selections made by the performer. Finally, mutual interpretation by the audience and the performer takes place during the performance. We are only concerned with the performer's interpretation of the audience (since the audience reaction is beyond our scope of direct manipulation!),

which is represented in the system as changes in the context model parameters (see section 3.2).

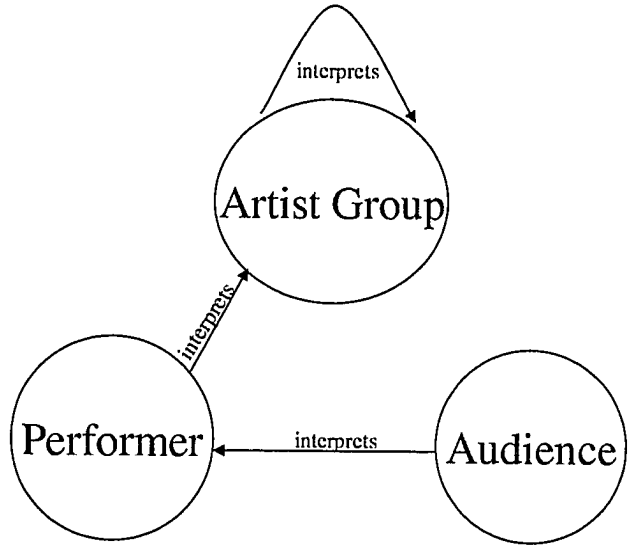

Figure 2: The interpretation triangle

Interpretation can occur at multiple levels. For example the Bible has a historical as well as a spiritual sense [Ricœur, 1969]. In creative practice, the phenomenon of interpretation is central in the artistic environment. For example, musicians interpret a composition, audience interpret artwork... It is not surprising therefore that we turn to a theory of interpretation to provide premises for our work. We decided to follow a text-based approach to the annotation of the data (vs. content based) using theories and techniques borrowed from the field of Natural Language Processing (NLP).

Interpretation is the consequence of ambiguity. In natural language, ambiguity can occur at many levels. At the syntactic level, a sentence may have several possible parses; at the semantic level, a word or sentence may have several meanings; at the discourse level, a text may be ambiguous due to the use of metaphors or referring expressions. In order to interpret a text linguistically, we need to select the correct

Table 1: Interpretation Scope

	Interpretation I	**Interpretation II**	**Interpretation III**
Scope:	artist group	performance	audience
Representation:	relations	dynamic selection	context parameters

syntactic parse and meaning of the words, sentences... This disambiguation may require the application of strict grammatical rules or may require the use of world knowledge or plain common sense. In art, while the process of interpretation seems more complex, it still involves using world knowledge to associate between the sensory cues and a certain meaning. However, the principal difference is that in the artistic context, it is more often that ambiguity is intentional and implicit.

It follows that a valid scientific representation of an artistic process or work would strive to retain this inherent quality of ambiguity rather than suppress it. We find this in contradiction with efforts in mainstream Natural Language Processing which often deal with technical texts where the focus is on disambiguating meaning [Jurafsky and Martin, 2000] (see [Manning and Schütze, 1999] for an example).

Another interesting aspect is that the Systemic Functional model is a probabilistic model that has been mainly applied to literal discourse and dialogue, however in our case we are applying it to a performance which is guided by poetry and metaphor. Indeed, as O'Toole remarks, art has non-communication functions such as exploration, discovery, and other imaginative functions [O'Toole, 1995]. This is one of the reasons why modeling Art has to account for subjective and sometimes irrational components.

1.5 An Example

In order to make our goal clear, let us look at an example inspired from the current production. Let us consider a presentation of about 45 minutes length. The presentation is about certain persons and places. The subjects of the presentation are loosely related. The knowledge base comprises one hundred visual text images, about two hundred video excerpts, two hundred audio samples including music, voice etc., about 1,500 images, embedded prior artistic knowledge and/or expertise, as well as rules for selection from the available media. Following is an example of what the presentation outline looks like:

Presentation Outline

I PK (5 min.)
II Cody (5 min.)
III Yan (5 min.)
IV Vera (5 min.)
V Brazil (5 min.)
VI Canada (5 min.)
VII China (5 min.)
VIII Discussion (10 min.)

When triggered for input, the system should start playing the material it predicts as the most appropriate or give the user a ranked list of choices to select from. The system may generate a mixed presentation of two or more selections (e.g. images with sound or separate windows one running a video and the other containing text). Triggers of the system could imply a search for specific data features, an indication

of changes in the context of the presentation or a request for applying visual or audio effects.

In Figure 3, a use case diagram describing the requirements of the system is shown. Two classes of users are depicted in the diagram, the presenter and the annotator. These could be the same person or different individuals. By annotator we refer to all those involved in the preparation phase of the presentation. The role of the annotator is to prefigure the presentation parameters including relevant semantic and technical data features, patterns of selections likely to recur, and any specific rules which do not concern the information retrieval model, and that need to be explicitly integrated in the system (heuristics).

1.6 Intended Applications and Users

Multimedia presentations are being used increasingly in different domains. We illustrate in this section some of the potential applications of the suggested framework as well as the classes of users who might interact with these applications.

1.6.1 Applications

Enterprise Websites offer more and more multimedia presentations and demos. Banks (e.g. http://www.scotiabank.com/cda/content/0,1608,CID4961_LIDen,00.html), car manufacturers (e.g. http:// www.mercedes-benz.com/), insurance, clothing, telecommunication companies and other type of business use multimedia presentations for the commercialization of lines of sophisticated products. In the education field, a presentation system can facilitate the task of professors using multimedia to enhance class presentations and who tend to select, according to the context, material from an

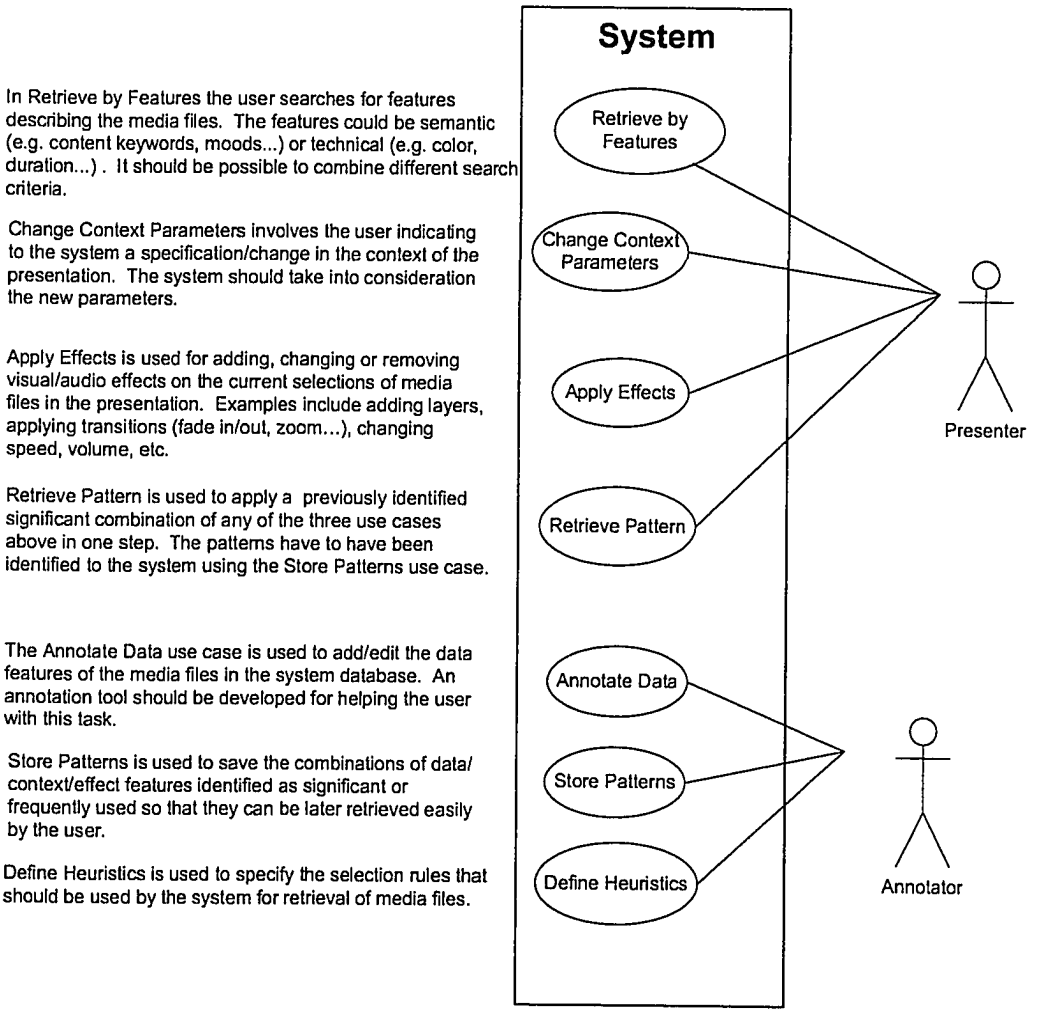

Figure 3: Use Case Diagram of our System

augmenting repository of information. In the culture domain, research has already been conducted in the area of automated museum guides with the goal to develop an adaptive multimedia commentary to the museum visitor [Not and Zancanaro, 2000] and [Not and Zancanaro, 1999], while we illustrate here an example of potential application in the arts and entertainment field.

1.6.2 Users

Direct users, those who physically interact with the application, as mentioned in the previous section, include the person who presents the material to others (e.g. a performer, a professor or a sales/marketing person) as well as those involved in the annotation process, either as conceptual decision makers or as annotators. Direct users could also be exploring by themselves material prepared by others as in the case of museum visitors who use automated museum guides and website users.

Indirect users do not interact physically or actively with the system, but only as spectators (e.g. students in class, audience of a performance). It is important to point out this distinction for the purposes of evaluation (see chapter 5).

1.7 Organization of the Thesis

In this first chapter we discussed our motivations for building a framework for adaptive multimedia presentations and introduced the theoretical background for the framework premises. Chapter 2 provides a non-exhaustive literature survey, covering general topics and specific projects in the area. In chapter 3 we lay out the proposed framework, and map onto it the example introduced in section 1.5. Chapter 4 describes an implementation effort which we used as a proof of concept, and discusses specific issues related to the architecture, platform, design and interface of multimedia applications. Chapter 5 deals with evaluation issues in multimedia information retrieval and in the last chapter we draw conclusions and suggest future directions of research.

Chapter 2

Literature Survey

Dieu! quelle importance ils accordent à penser tous la même chose.
Jean-Paul Sartre, La Nausée

There is a considerable amount of research in the area of Multimedia Information Retrieval and Presentation. Much of this work is driven by the Internet explosion. Content-driven multimedia retrieval appears to be the main focus. Other areas include knowledge management, modeling, database and user interface issues. In this chapter we discuss the current trends in Multimedia Information Retrieval and Presentation, then focus on some specific projects in the area.

2.1 Frameworks/Models

Modeling attempts in the multimedia domain are often task, domain, process or media dependant. Due to high complexity, it is necessary for any framework/model to strike a balance between generality and applicability. Examples of models include the context, user and interaction modeling for museum context in the HIPS

project [Marti *et al.*, 1999] (see section 2.6.1 for further discussion). A Visual information annotation framework is described in [Jaimes and Shih-Fu, 2000] (see section 3.1.1).

The MATN (Multimedia Augmented Transition Network) [Chen *et al.*, 2002] is particularly interesting since it proposes a general model for live interactive RTSP (Real-Time Streaming Protocol) presentations. It is worth noting that the description of the model by the authors as a "semantic" model refers to the semantics of interaction and presentation processes rather than the semantics of the content. Specifically, MATN is used to model actions (e.g. Rewind, Play, Pause...), temporal relations and synchronization control (e.g. concurrent, optional, alternative). The authors of the model claim its novelty lies in combining the modeling of user interactions, loops and embedded presentations.

While the MATN model allows for a certain adaptability through its support for user selections, this adaptability is severely limited, since there is no information retrieval component in the model. This implies that all alternatives must be explicitly identified by the designer of the presentation beforehand; an impractical option in the case of huge repositories with numerous alternative paths. Furthermore, it is not possible in this model to reason about the semantic content of the media and take actions consequently. As for context representation, it is limited to certain time variables such as *start of presentation* and *current time*, which again does not permit the model to learn about users preferences in such a way to make intelligent choices in response to users selections.

2.2 Content-based vs. Text-based

Content-based retrieval refers to the automatic identification of relevant semantic and non-semantic content of the subject medium and querying based on content similarity. Content-based retrieval is regarded as the ideal in multimedia research. In this approach pattern recognition techniques are usually employed in order to automatically identify content and understand its semantics. State-of-the-art in pattern recognition allows modeling closed domains with some success. Optical character recognition is a good example of successful pattern recognition of textual data. While relative success has been achieved in speech recognition, sound classification [Blum *et al.*, 1997], as well as color and texture identification as in [Flickner *et al.*, 1997], complete image and audio understanding, including the extraction of complete semantic descriptions and interrelations, is unlikely to be attained with the currently available technologies.

On the other hand, text-based retrieval uses textual metadata to convey the semantic and technical information of the media file. Text description is annotated either manually or using semi-automated tools as discussed in section 2.4. Querying is then performed within traditional text information retrieval models as those described in section 3.7.

While content-based retrieval reduces the extent of human involvement in the annotation process, it could also prove inflexible when subjective judgment needs to be applied. In order to review and/or override the automatic annotation, this later needs to be in human-readable and modifiable format, which is not the case for example of raw data, histograms and other frequently-used content-based annotation. To resolve this issue, a trade-off between text-based and content-based

annotation could be a mapping scheme which translates raw data into equivalent semantic notation (For example VisualSEEK uses color set transformation of histograms [Smith and Chang, 1997]). This could be especially desirable in closed-domain tasks which could utilize domain ontologies for mapping.

2.3 Context-based IR in Search Engines

Recent directions in text Information Retrieval research show a shift in focus from content-based approaches through user modeling, and finally to context modeling. The quest of search engines is to arrive at a more precise and complete result set of relevant information to meet the user's query. This is more critical in the case of multimedia retrieval. Whether content-based or text-based, queries for multimedia databases tend to be more ambiguous in nature than queries for text databases. Short ambiguous queries return potentially low-precision results. Better performance for search engines starts at the query formulation phase. Precisely understanding what the user is searching for would certainly improve performance.

In order to overcome the obstacle of understanding the query, some search engines experimented with building user models, based on roles and professional interests. However no significant improvement was observed [Goren-Bar *et al.*, 2001]. For this reason, researchers have started experimenting with context-based information retrieval. Context refers to information which can not be deduced from the query terms, but forms part of the environment of the query, such as the goal of the query, and in a broad sense it encompasses the user who makes the query. *Prism* [Leake and Scherle, 2001] is such a search engine which attempts to extract contextual information using the Watson method [Budzik and Hammond, 2000] to monitor the

user's activities in standard applications like word-processing. The Watson method makes use of style characteristics of words such as emphasized text, as well as the location of words in the document being authored as contextual indications of the importance of these words for queries. It then uses heuristics and traditional information retrieval methods (TF-IDF) (see section 3.7) to infer the selection of specialized search engines to which it directs the user's query. Results from Prism suggest that using contextual information for this task can improve the retrieved results' usefulness.

ACQUIRE (Adaptive Constraint-based Query Interface) [Huang *et al.*, 2001] is yet another project making use of the interaction with the user to dynamically build a meta-search engine interface. Interactions with the user can be considered contextual information. Indeed in our implementation we allow the user to control the presentation through constraints on features and contextual fields.

2.4 Annotation and Tools

Manual annotation of multimedia data is a very time consuming and tedious task. As mentioned is section 2.2, some objective features can be annotated automatically by using techniques developed in pattern recognition and in text analysis. It is also possible to use semi-automated and computer-aided annotation tools.

A particularly interesting effort in this domain is the framework and manual annotation tool proposed in [Jelmini and Marchand-Maillet, 2003] as an extension to the Dublin Core Metadata Initiative which promotes the development of interoperable metadata standards (http://dublincore.org). The model, described in ontology language DAML+OIL, can be extended using any specialized ontologies according to

the domain.

2.5 The User-Interface in Multimedia Systems

Multimedia Information retrieval interfaces can be quite complex depending on the required functionality (see section 4.6). Interface features can be categorized according to functionality in two groups: On one hand presentation features supporting continuous play including a control panel are similar in many aspects to presentation software such as Macromedia Flash and MS-PowerPoint, with the exception that some of the design-environment features such as slide transitions in MS-PowerPoint or Alpha and Color selection in Macromedia Flash could be needed as interface elements. On the other hand, retrieval features including querying, browsing and relevance feedback belong to the realm of search engines.

Cluster-based representations for returned documents are suggested by Au et al. in [Au *et al.*, 2000] and Carey et al. [Carey *et al.*, 2003], who designed several cluster-based visualization interfaces for text document search engines using keyword generation. Sammon-mapping, Tree-map and Radial visualization are experimented with. In these experiments, returned documents are clustered and labeled by subject keywords. In Sammon-mapping, each cluster is represented by a circle on the screen, whose color and size determine the size of the cluster, while distance between clusters indicates their relative similarity. As a user browses through a cluster, she can see the description of documents and their URLs in a bottom panel. Keyword refinement is possible within and across clusters. This approach allows moving between browsing and searching. The tree-map algorithm points out the second-order cluster structure including the frequencies of the related keywords used in building the clusters, while

the radial visualization technique allows the user to interactively build the clusters.

2.6 Research Projects

In addition to the related general research domains discussed above, specific projects have dealt with issues similar to those we handled in our project. Following is a brief comparative description of such projects.

2.6.1 The MacroNode Approach

In the HIPS (Hyper-Interaction Within Physical Space) project, a portable electronic museum guide developed by Not and Zanacanaro ([Not and Zancanaro, 2000] and [Not and Zancanaro, 1999]), 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.

The end-user, 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. For example, the visitor can be in front of the object described, either in close proximity or not, or alternatively not in front of the object. A composer-engine uses rules to build a presentation targeting both coherence and cohesion.

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 (see section 1.3). (In a later project by Zancanaro et al.

[Zancanaro *et al.*, 2003], the utilization of RST relations is extended to producing video like effects from still images, driven by the audio documentarya). A closed-set ontology is employed as well as an annotation tool. We account for an information retrieval model in the framework, while the MacroNode approach is only concerned with the presentation aspect.

As can be seen the MacroNode formalism bears resemblance to the framework suggested in our project, with differences due to its task-based and domain-specific model. While it is possible to augment the domain model using other ontologies, this would not account for context parameters. The concept of contextual presentation is limited here to physical location and subject as context, while in a more general model other factors such as time, history and audience should be accounted for. Only self-presentations are relevant in museum guides, hence relationships between presenter and audience is not considered. Also the presentation building blocks, which consists of fragments of phrases, are relatively closely knit and can only provide for minimum variations.

In our framework, no specific size of a data component is recommended, since time is modeled as a context element, although smaller data tend to allow more flexibility in content selection. Furthermore, we include such components as the domain model and the user model in our framework, whereas they are considered external resources in the MacroNode approach. Also we do not attempt planning yet, instead responding to changes in the context by applying appropriate heuristics. As we gain more expertise in the field, we hope to acquire the knowledge required for planning.

2.6.2 Planning Cinematographic Structure of Animations

Kevin Kennedy and Robert E. Mercer [Kennedy and Mercer, 2002] developed a communicative act planner using techniques from Rhetorical Structure Theory (RST) (see section 1.3). 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. The planner constructs a tree from the rhetorical relations to convey a coherent outcome of the animator's plans. In this respect, similar to the MacroNode approach [Not and Zancanaro, 2000], the focus is on small building blocks.

This project is similar to ours in that it does not attempt to create new animations but generates altered ones. However our tool is intended to be used both during the preparation phase as well as during the actual presentation. Also notable is the necessity to adapt RST relations to the visual domain. These relations are used here to generate effects for communicating to the viewer. The planner constructs a tree from the rhetorical relations to convey a coherent outcome of the animator's plans.

The knowledge used in this project to associate techniques with thematic and emotional effects was acquired from a film studies textbook. For example, classification of lighting according to energy is associated with certain emotions from which the user makes selections. While animation is an established field with well-defined practices, we believe that in the case of generic multimedia presentations, where animation is only one medium, and for which no standard practices have been developed,

it is more convenient to provide the user with a flexible way to experiment with techniques. However, we did include in our framework a place-holder for a knowledge base in anticipation of developing the required expertise.

Besides general and common research areas pertaining to Multimedia Information Retrieval and Presentation, we presented specific projects with certain adaptability to user and context parameters. In the next chapter, we introduce our framework for adaptive multimedia presentations, and in the following chapter, we present a specific implementation of the framework as a proof of concept.

Chapter 3

The Framework

On ne peut invoquer la nécessité de l'ordre pour imposer des volontés.
Albert Camus, Actuelles.

In this chapter we describe the proposed framework for modeling adaptive multimedia presentations. Figure 5 is an illustration of the framework components. In this figure Static refers to those elements of the framework that do not contribute directly to the adaptive potential of the framework and could be fixed for different presentations. The dynamic elements are those responsible for the adaptive aspect of the framework.

The framework is intended to be general enough to be used in different contexts. This means that not every presentation will need to satisfy all the components. In general, more complex requirements for a presentation will require a more elaborate utilization of the framework. Implementation issues, such as the architecture and the user interface are deliberately excluded from the framework. They will be addressed separately in chapter 4, in which we present a prototype system as a proof of concept. We would also like to emphasize that we present in this chapter hypothetical

Figure 4: The Framework Glossary of Terms

Data model: The logical representation of the data.
Context model: A representation of the environment.
Feature relations: Relations between the component features.
Information retrieval model: The retrieval component.
Heuristics: Rules for mapping *Context* into *Content*.
Patterns: Interesting combinations of features for fast retrieval.
Effects: Visual and Audio effects applied to the selections.
Knowledge base: Permanent expertise in the field.

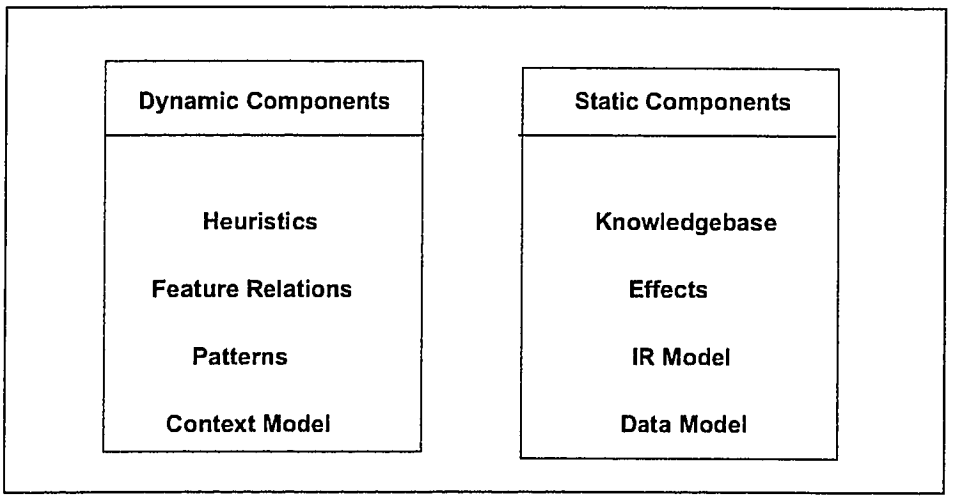

Figure 5: Framework for Adaptive Multimedia Presentations

alternatives as design solutions, which have not been implemented. The actual implementation includes the data model, the information retrieval model and limited areas of the context model, selection heuristics and effects. The mini-glossary that follows, describes briefly the components of the framework. We will introduce these components in more detail in the following sections, illustrating with examples from the current project where possible.

3.1 The Data Model

The data model consists of the data files in any format that can be supported by the visualization software, (in our case Flash-MX which supports formats including mp3, mpg, swf, html...) and the annotation of these with technical and semantic features as well as relational characteristics. As proposed by Prabhakaran [Prabhakaran, 1997], multimedia objects can be modeled as a general class with specialized classes for each type of media.

Meta-data which describes semantic features of the media files is constant across media types. These include *Keywords* describing the semantic content and the Mood of the selection. Besides available ontologies for classification (see section 3.1.1), categorization can be achieved according to generic criteria, for example at the *General*, *Abstract* and *Specific* levels as applicable. By *General* we refer to a class of objects with physical presence like human, chair, dog. *Abstract* is an idea or concept without a physical presence such as hunger, war, sleep, whereas *Specific* is used for identifiable named entities (e.g. PK, Cody, The Oak-tree, World-War II). This general classification could make the data more reusable in other contexts and even by other presenters if desired.

Each type of media is also annotated according to its specific characteristics. For example, images are annotated with Color and Texture, while sounds features include Type (music/spoken/electro-acoustic), Duration, Dynamics and Pace and video is annotated with Frames/Second. Finally this model is extensible through the use of any relevant ontology, according to the specific features of the data and desired presentations. In the context of the current project, it was desirable to 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).

When deciding on the technical features to use for the different media, we did this in consultation with the artists involved in the production of the specific medium and after review of the literature. Since at this phase of the project all annotation is done manually, a trade-off between the features according to usefulness was necessary. Thus, some features which would normally be annotated were excluded. This will be indicated where appropriate. It should also be noted that certain technical features, for example *Colors* in visual annotation, were annotated in a subjective manner due to the artistic nature of the application (see section 3.1.3).

3.1.1 Ontology/Taxonomy of the Domain

Ontologies are hierarchical classifications of concepts and their relations within the data model. Using ontologies in information retrieval limits the scope of the search and establishes implicit relations between these concepts. General Information Retrieval systems use ontologies for classification purposes. In closed task-based or domain-specific models, developing ontologies can simplify the solution and permit the implementation of more complex processing. For example, heuristics can be developed, customized application interfaces can be designed, and user profiles modeled based on ontologies. Ontologies are linguistic by nature: building ontologies employs the domain taxonomy/terminology thus eliminating ambiguity. Ontologies are also useful for building semi-automated annotation tools. By giving the user a limited set of choices, errors in data entry can be reduced, and a user-friendly interface would save substantial time for the annotator.

Table 2: Example of the annotation of three data files

document	physicality	emotion	mental space	media	colors	keywords
DSCN01	body (medium)	puzzled	reality	image	b/w	general: chair/snow
Text02	landscape (big)	thoughtful	dream	text	black	specific: Cody
DSCN05	page (small)	puzzled	metamorphic	image	blue	abstract: automation

Research on indexing visual data spans the fields of psychology, and information science. Work by Panofsky [Panofsky, 1962] and later Shatford [Layne, 1986] classifies visual information into *pre-iconographic* and *iconographic* and *iconological*, where the first refers to information pertaining to the objects present, while *iconographic* refers to stories and conventional semantics and *iconological* deals with the interpretation of the image. In our model, *pre-iconographic* corresponds roughly to *General* and *Specific*, *iconographic* to *Abstract*, while *iconological* is dealt with through the presentation itself (e.g. order of selection, heuristics, moods...). At this phase of the project, all annotations have been done manually. Table 2 shows an example of some data annotations according to the features explained above.

3.1.2 Text

To define the borderline of text as a medium is a tricky task. Audio data could contain spoken text, images may contain text fragments (see Figure 6 for example), even complete poems could be laid out in a visual way like many of French poet Apollinaire's poems (see Figure 7), while 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). As such, text is the most researched medium in the field of information retrieval.

Many text search engines exist today which commonly employ prior indexing

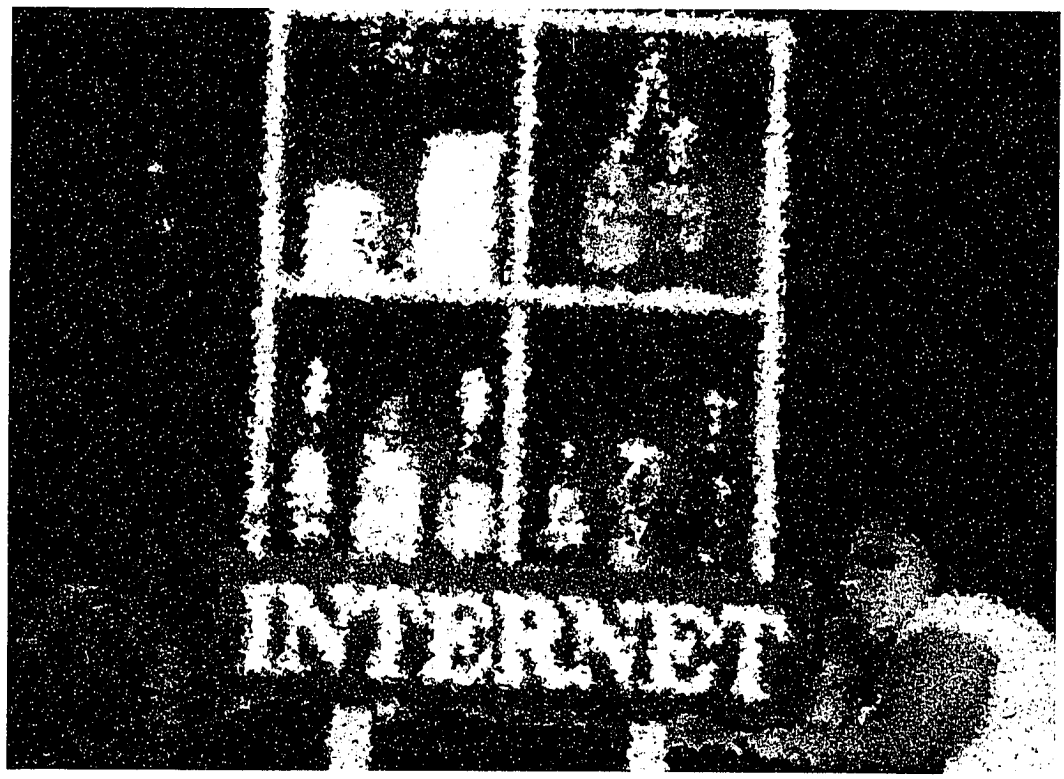

Figure 6: Example of Text Fragments within an Image

of keywords. In the context of multimedia presentations, text could have a duration property and a spatial dimension. These attributes could be handled by the heuristics or left to the user's discretion.

3.1.3 Image

Still images are digital graphics containing drawings, paintings, photographic images, text or any combination of these. Raw image data consists of usually compressed bitmap pixel data. The most common format is JPEG. Technical features commonly used for annotating graphics include colors, texture, dimensions and file format. In this project file formats are limited to JPEG and dimensions are also limited to a

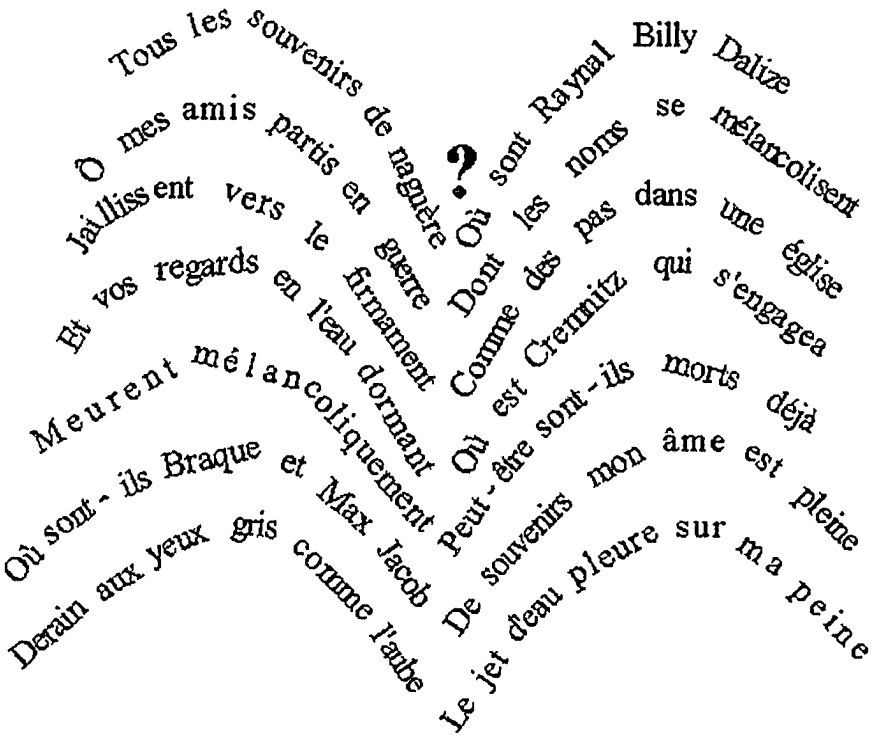

Figure 7: Visual Poem by Apollinaire

maximum size, while texture was deemed irrelevant so these features were excluded.

Multiple color annotations were permitted. Colors were listed in order of significance; however this aspect was not utilized in queries. In content-driven retrieval, color histograms or sets might be used. Histograms represent color percentages, while sets use thresholds to include only significant colors and are thus not as expensive computationally [Smith and Chang, 1997].

3.1.4 Moving Images

Moving images include videos and animations. This category does not include sequences of still images which are handled using relations, heuristics, patterns and the display logic. We use atomic excerpts consisting generally of a few seconds to

Figure 8: Example of Video Annotation

History of China
Class: Video
Duration: 4 min.
Pace: moderate
Main goal: Inform (Background)
Other goals: Solicit
Keywords: {China, History}
Main Effect: Nostalgia
Other Effects: High Spirits
Color: Red, Yellow
Author: {PK, 2003}

two minutes. This roughly corresponds to the definition by [Carrer *et al.*, 1998] of a Scene. A Scene is a collection of contiguous logically related shots, while shots are contiguous frames with common content. Sequences which are the higher level in this hierarchy are not considered as units, but are dealt with through relations. This was an optimum solution to avoid segmentation process, and since shots were not found a relevant unit.

Together with Audio this category has a time dimension. It is 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. A typical video might be annotated as in figure 8.

3.1.5 Audio

Audio can be music, speech, or other sound data (e.g. Electro Acoustic, noise etc.). Other than this classification, temporal features have to be indicated for audio, like duration. In the case of music we indicate pace (tempo) using qualitative attributes

Table 3: Example of the annotation of audio files

document	type	emotion	mental space	pace	dur.	keywords
PK reads poem	text	thoughtful	dream	slow	60s	female
Osama poem reverb	elec. acoustic	thoughtful	dream	slow	45s	Osama

(fast/medium/slow), as well as the type of music (melodic, harmonic, percussive, gestural).

Thorough annotation of technical musical characteristics - in the case of music pieces - was one of the options investigated. It was found potentially useful had it not been for time and resource limitation. This includes, for example, the quality or timbre of the sound and the type of music (percussive, melodic, harmonic).

Akin to text data, speech carries information in natural language. Speech recognition technology is often used for pre-processing of speech data. However, this is beyond the scope of our current research, since it belongs to content-based retrieval.

3.1.6 Relations

Relations between the media files are also annotated. As mentioned earlier we use modified RST-like relations. There are two purposes for these relations. The first is 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 syncronization of two or more selections. The other purpose of using Relations is to support a relational navigation map which could be used to traverse selections according to their sensory and/or semantic links.

As in the MacroNode approach [Not and Zancanaro, 2000], multiple relations can be represented. This allows for web rather than tree structures, which is customary

Table 4: Example of the annotation of relation for one image file

document	requires	precedes	phonetic
DSCN0001	DSCN0005, V0002	none	s1.mp3
DSCN0005	none	V0002	none

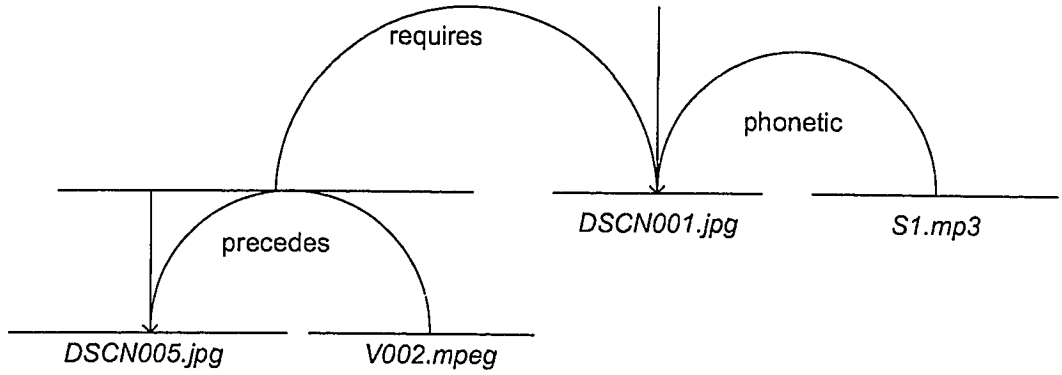

Figure 9: RST-like relations representation

- though not a requirement - even in the case of text structure [Marcu, 2000]. Since RST relations are semantic in nature, we had to augment these with new relations, which describe temporal constraints (follow, precede, simultaneous) and others that express pure sensory associations (phonetic, visual).

Table 2 shows an example of the annotation of these relations for the image files DSCN0001 and DSCN0005. In this example, the relation "requires" implies that files, image DSCN0005 and video V0002, are required to be played before image file DSCN0001 (although not necessary in that order). The relation "precedes" enforces the playing of file V0002 before DSC0005.

Figure 9 shows the representation of the relations between the media files in RST format.

3.2 The Context Model

As mentioned in Section 1.3 we attempt to draw parallels between *Context* in Systemic Functional model and multimedia presentation context. In the Systemic Functional model, *Context* refers to the environment, or any relevant features of the situation. For language, Halliday defines it as the following [Halliday and Hasan, 1989]:

Field: The social action that is taking place.

Tenor: The participants, their status, roles and relationships.

Mode: The channel of discourse (written/spoken/both) and the rhetorical mode.

For the purpose of multimedia presentations, we define context to include several interdependent features : the outline, time, space, presenter, audience, medium, rhetorical mode, mood, and history. As will be described in the next sections, some of these variables correspond to *Context* in SF, while others are either implicit in language or particular to multimedia. Although these context variables have been identified in our model, they certainly do not represent a closed set; a more refined framework could of course use several other such features. Figure 3.2 shows a snapshot representation of the context of a presentation. We give here a brief description of the context variables.

3.2.1 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. This information can be used to select the most relevant subject material. The user is able to change the subject through the interface of the system, triggering a system response in the form

Current Section: PK
Presenter: PK
Mode: Descriptive
Medium: Video/Image/Voice
Audience: Artistic
Time: 180s.
Space: Inside
Follows: None
History: pk-speech.mp3
Moods: Dramatic

Figure 10: Representation of the Presentation Context

of new material related to the current subject. Examples of the outline could be the different halls, or a chronological order in the context of a historical museum presentation, or the species in a natural history museum. Outlines could be more complex and contain overlapping sections.

The outline can be ordered or unordered and should support time constraints as needed. Relevant keywords can be used as a representation of the topic, by applying the same ontologies used for data annotation (see section 3.1.1). If desired a weighting scheme should be used to indicate the relevance, or relative importance of each keyword in a section. Similarly rhetorical modes (see section 3.2.6) and moods (see section 3.2.7) for each section, whenever possible, should be fed to the system. This information will be useful in discovering and experimenting with the system's heuristics (see section 3.4). The following frame shows a possible representation of one of the sections from the outline presented in 1.5.

Section: PK
Duration: 10 min.
Main goal: Introduce ()
Other goals: Inform (background)
Main Effect: Intrigue
Other Effects: {}
Prepares: China
Recaps: Cody
Follows: None
Keywords: {PK, Cody, China, places, red}
Requires: pk-speech.mp3
Tempo: moderate

Figure 11: Representation of a Section from the Outline

3.2.2 Time and Space

Time represents the timeline of the presentation. Timing 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...). It should be possible to relate Time and the Outline of the presentation and to dynamically change this relation.

The physical size of the space as well as its placement (interior/exterior) provide hints to the appropriate type of media to play. Presets can be determined to handle different space configurations.

By *virtual space* we refer to the spatial layout on the screen. This is relevant when

multiple objects are presented on the screen simultaneously or when one object does not occupy the full viewing space.

3.2.3 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 (see section 3.1.1) to avoid an extra step of matching terms. Other user profiling techniques include registering the users' requests to determine their interests.

3.2.4 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, which is the case also of presenting from a position of authority as opposed to a peer-to-peer presentation.

Employing stereotypes has become a common practice for modeling anonymous audience, especially in web-based applications which service a significant number of uses with varying characteristics and interests.

3.2.5 Media

The media of communication being used at a given time in the presentation is also important. Possible media include video, audio, animation, image, text and combinations of these, whether simultaneous or overlaid. This 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 knowledge of the currently playing media types as a context parameter is essential for heuristics. For example, a heuristic could decide whether or not to interrupt the current selection based on this information.

3.2.6 Rhetorical Mode

The rhetorical mode is the communication strategy used at a given moment in the presentation. Different rhetorical modes are intended to affect the audience in particular different ways. Examples of rhetorical modes given by Halliday include persuasive, expository and didactic [Halliday and Hasan, 1989]. These criteria correspond to some of the perlocutionary effects as described in Speech Act Theory [Searle *et al.*, 1980], while the emphasis is on the effect rather than the act. In the context of our framework, it seems more convenient to take into consideration effects rather than acts, due to the complex and imprecise nature of visual and other non-speech acts. 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 .

The rhetorical mode is potentially related to the type of audience and their nature of response: the same presentation could be presented for example in process analysis mode to scientific audience and narrative mode to artistic audience. The relative timing in the presentation might also affect the choice of rhetorical mode. Different segments of the presentation could employ different rhetorical modes, while the presentation as a whole could belong to a single mode. By informing the system of the desired rhetorical mode at a given time, appropriate heuristics (see section 3.4) could be applied and generation patterns (see section 3.6) selected. For example, the rhythm applied for Illustration could be different from that applied for Definition. Also retrieved results could be ordered in an alternating fashion to emphasize Comparison/Contrast mode.

3.2.7 Moods

The emotional feel of the presentation or its mood contributes to maintaining a coherent context. For example, Kennedy and Mercer have applied visual effect to alter the emotional predisposition of the viewer for animations [Kennedy and Mercer, 2002]. 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. This cannot be done without certain subjectivity, and hence could be also specified per user profile. The definition of Moods themselves has to be qualitative and may be comparative (e.g. happier, happy, neutral, sad, sadder).

Moods are directly affected by the different elements of the media. 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.

3.2.8 History

A record of selections already played could be kept in history and 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 (see section 6.4).

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 use 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

Heuristics are rules to be applied in specific situations. Heuristics are not error-proof and some could contradict others, requiring further heuristics to apply the most appropriate.

3.4.1 The Goal of Heuristics

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 at a given moment is mapped into specific selections. This context includes the performer and the audience model, space, time, selections already played, in addition to any explicit triggers such as a request for different moods or artistic patterns and techniques.

Experimenting with the selection heuristics will allow us to refine them and will provide the performer or an external observer the ability to examine the artistic cognitive process and to discover artistic ideas, patterns and techniques, specific to each performer, which can then be fed into the data and context models to customize the production of the presentation to a specific performers style.

3.4.2 Heuristic Patterns

Prediction heuristics will then be added in the form of rules for guiding the selection of media to play. These might include for example giving more importance for the time restrictions, trying to play first a selection that fits in the mood of the current section or searching related content first, insuring that all themes are covered adequately and

avoiding unnecessary repetition of themes and goals, or attempting smooth transitions between different sections by playing a boundary selection.

To find the relevant heuristics, experiments should be conducted during rehearsals through variations of media combinations, selections duration and content, intended effects, and tempo. The experimentation can result in the identification of features for high-level classification patterns of the knowledge base and situation-related heuristics. For example, classification patterns might be related to genre (impressionism, expressionism, realism, abstract, surrealism...) or simply describe links between certain features like tempo and intended effects, while the heuristics might depend on the parameters changed by the user.

These patterns can then be used for querying the knowledge base. The system will attempt to reproduce the pattern requested by the user by searching the pattern space as well as applying the appropriate heuristics in the given situation.

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. For example if the performer decides to introduce the new theme of "China" after the introduction instead of at the end (see Section 1.5), the system will apply the heuristics which might produce a linear equation of weights and factors in the form (factor1×weight+factor2×weight+...). An example might

be Remaining_time×.2 + Mood×.3 + key_words×.3 +Main_goal×.2.

The system will then use this equation to search for the most appropriate selection. If the user requires a change in the tempo, the system might respond by varying the speed of displaying the images.

It might be possible to provide for ways that the user/performer can use to dynamically override the system's selection, to ask the system, by indicating a different preference through relevance feedback, to try another time. However this could cause a problem of interruption to the flow of the presentation, especially if the response time of the user interface is not adequate to real-time tasks.

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 as in Internet Explorer 4.0 and later versions. However, including them in the run-time interface in an accessible manner, would allow 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. A discussion of the usage and semiotics of transitions and effects in cinema can be

found in [Metz, 1968]. [Zancanaro *et al.*, 2003] implemented cinematic techniques such as cut, fade-in and fade-out in a multimedia museum guide using Macromedia Flash.

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/heuristics. Similarly to heuristics, generation patterns are discovered while experimenting with the system during the rehearsal/preparation phase of the presentation. Once identified, it is possible to retrieve them explicitly during the presentation by including them in the interface. 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 simplifies the presenter's task by giving a shortcut to a goal otherwise difficult to achieve 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

As mentioned earlier, any 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. In this model documents are represented as vectors

of relevant term weights and ranked according to their distance from the query vector. One common scheme for assigning weights to terms takes into consideration the term frequency and the inverse document frequency (TF*IDF). Common terms across documents are thus considered less important and assigned less weight. Other models include the Boolean and Probabilistic models. For a thorough discussion of information retrieval and the different models see [Baeza-Yates and Ribeiro-Neto, 1999].

3.8 The Knowledge Base

A knowledge base consists of data and rules in machine readable format. 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 and heuristics. The knowledge base might also include ontological hierarchies of intentions, strategies, meanings, effects and rhetorical relations. The knowledge base might be annotated to reflect communicative goals (e.g. introduce, convince, demonstrate, recap, summarize), relationships between documents, and features like effects and relevant content keywords.

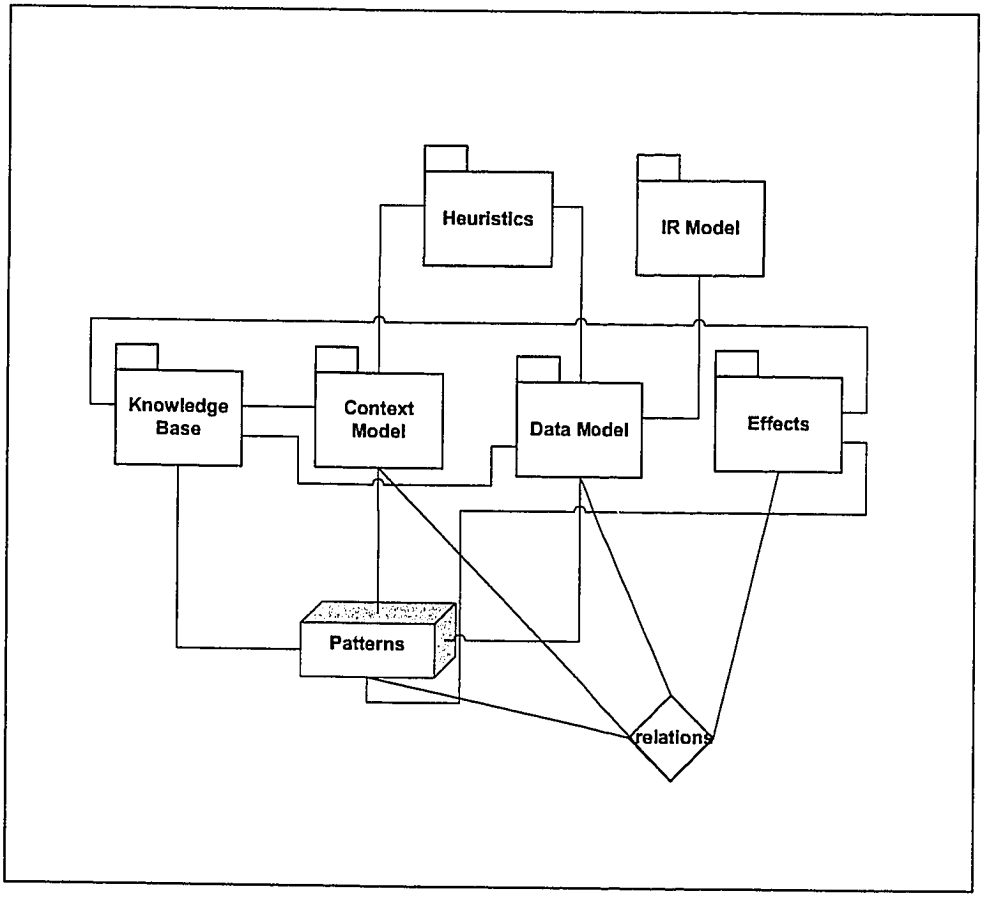

Figure 12: Framework Relationships

3.9 Relationships between the framework elements

A complex network of potential relationships exists between the components of the described framework. Figure 12 depicts the relationships as described in the previous sections. Following is a brief recapitulation of the significance of these relations:

- Heuristics map contextual conditions as described in the Context Model to actual content represented in the Data Model.
- The IR model defines the way selection criteria are applied to retrieve data from the Data Model.

- Patterns are meaningful recurring conditions which could combine elements from the Context model, the data Model and Effects.

- Relations can establish preferential links between Attributes (Features) in the Data Model, the Context Model, Effects and Patterns of more subjective and temporary nature than Heuristics which are more constant rules represented in programming.

- The Knowledge base includes expertise in the field of Multimedia such as Heuristics, Effects, Contextual conditions such as rhetorical modes and rules to apply this expertise on the data model.

In this chapter we have introduced a conceptual framework which we hope can accommodate a wide range of adaptive multimedia presentations. In the next chapter, we present a prototype system partially implementing this framework.

Chapter 4

Implementation

Reality is merely an illusion, albeit a very persistent one. Albert Einstein.

A prototype of the proposed framework has been constructed. It was developed using a three-tier software architecture on flash/java/MYSQL platforms. In this chapter we describe the implementation and discuss some of the decisions underlying the architecture and platform choices of the system. Section 4.1 provides an overview of the system life-cycle, while section 4.2 describes the software architecture. In section 4.3 we list and compare some of the platform choices considered. Section 4.5 illustrates the design of the system and in section 4.6 we discuss the interface.

4.1 Life-Cycle

As Arts and Science seem to diverge with regard to purpose and approach, comparing artistic and scientific methodology might seem far-fetched at first sight. Not so if we consider them from a phasic point of view. The life-cycle of the system is iterative and reveals a striking similarity to that of the artistic performance. In fact this analogy is

not coincidental. The artistic and scientific processes naturally converge at the same point: The prototype-rehearsal. Since the system is intended to be used for real-time performance, high fidelity prototypes are used to review both the functionality of the system and the content of the performance. A high-fidelity prototype is one which resembles as much as possible the final product, functionality and interface-wise, with specific attention to interactivity. Decision makers and direct users are involved in this process. In earlier stages of system design, prototypes are used to refine requirements, while in later stages they help in establishing feature relations and in discovering retrieval heuristics and patterns for recall. The principal processes of a cycle include:

1. Analysis of the provided media in order to extract information and identify correlations between the different excerpts.

2. Deciding on a representation of the media and correlations.

3. Manual indexing of the material according to the chosen model for the knowledge base.

4. Developing a presentation model also based on the knowledge base model.

5. Developing an interaction (dialogue) model probably based on communicative acts to translate the user's wishes to the system.

6. Identifying the system's prediction heuristics.

7. Designing a user-interface to manage the human-computer dialogue and the presentation model.

4.2 Architecture

Long-term experimental goals, the volatile nature of requirements and other practical considerations have influenced the three-layers architecture choice for the system. Experimentation and changing requirements entail possible changes in platform and technologies. The separation of the presentation from the model and the business logic allows for the substitution of any of these layers at minimum cost. This is important given the versatility of technologies in the relatively recent multimedia field (see section 4.3).

Figure 13 illustrates the three tiers comprising the architecture of the system. The model tier consists of the data and annotations, relations and retrieval patterns. Business logic including operations on the database, heuristics and status information make for the middle tier, while the presentation (view) tier has the user interface and interaction elements.

Communication between the presentation and business logic tiers is done using XML socket, while operations on the database are performed in SQL (Structured Query Language). This architecture allows maintainability, scalability, and future expansion of the system to use multiple views for different tasks, users or contexts.

4.3 Platform Choices

Platform choice was in part dictated by the three-tier architecture preference (see section 4.2) and object-oriented design. In order to insure a loosely-coupled architecture, one with least inter-dependence between the layers, ideally each layer should be implemented on a different platform. Other factors involved in the selection of

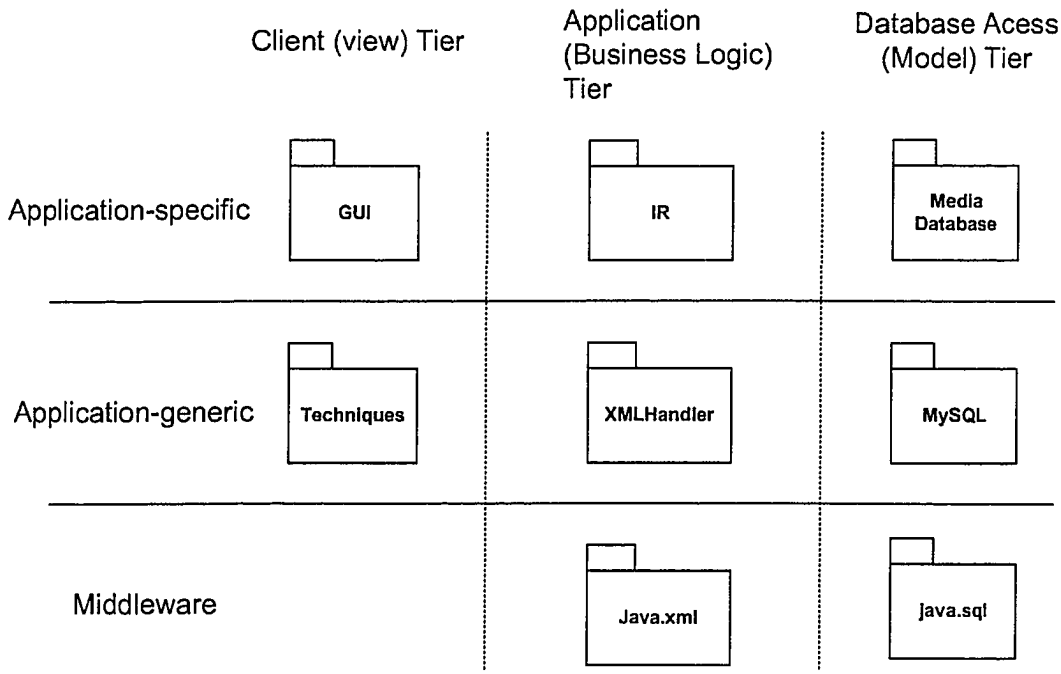

Figure 13: Architecture of the System

the platform include functional requirements such as persistence (uninterrupted play of media files), support for specific media types, (namely mp3, mpeg, and SWF), as well as practical considerations such as the software cost and the use of mostly non-proprietary software. Following is a presentation and comparison of some of the platforms considered:

1- Java Media Framework (JMF) + Java + Database engine: Java Media Framework (JMF version 2.1.1) is a recent promising framework developed by SUN for manipulating multimedia. Based entirely on Java technology, JMF supports capture, playback and streaming of audio, video and other multimedia. It follows that JMF integrates very well with Java. Careful object-oriented design would ensure the separation of the presentation from the application logic. However, there are a few

disadvantages of JMF: Being very recent technology raises question about its reliability and maintainability. Another disadvantage is its lack of support for various media formats including Macromedia SWF (shockwave flash) file format (older version of JMF supported SWF v2.0), MPEG, as well as MP3 (due to licensing requirements). Finally its relative complexity compared to Macromedia Flash can potentially make it harder for non-IT team members, like artists, from being involved in the interface design.

2- All Macromedia technology solution: A straightforward solution might seem to use an integrated Macromedia solution for both front-end and back-end operations. Macromedia Flash-MX does not support data connectivity. Instead ColdFusion, JRun, Flash Remoting or Flash Communication Server should be used for server side operations. The benefits of this solution include that Macromedia server-side solutions integrate smoothly with flash player for client-side presentation, also flash client support for MP3, MPEG in addition to its proprietary format SWF. However, it is quite likely that this solution leads to tightly-coupled layers, due to platform dependencies. Another concern that Macromedia Server-side technologies are proprietary software with a significant cost; an obvious inconvenience in an experimental system where platform change is envisageable.

3- Macromedia Flash + middleware + database engine: This solution offers the advantage of separation of tiers by platform. Some of the considerations for choosing the application-tier platform cost, complexity vs. familiarity, extensibility and object-oriented capabilities. In this implementation, MySQL, a free and reliable database engine, was chosen for back-end operations. For the middleware, we

preferred Java 2 standard Edition(J2SE) with its support for database connectivity(JDBC), SQL(java.sql) and extensions for XML support(org.xml,org.jdom).

4.4 Strategy

In the context of this project the problem of the implementation of the framework can be divided into three main tasks:

- Modeling the presentation and the audience.
- Discovering the appropriate heuristics for the selection of material.
- Projecting the human-computer interaction as a genuine component of the artistic experience.

The system needs to be prefed the information in this outline at "compile-time". This will be done in terms of the top-level communicative strategy of the presentation (i.e. communicative purposes of each section/subsection and how the sections relate to each other). It then serves during the preparation and rehearsals in discovering the prediction heuristics of the system.
In addition the system needs the following information:

- Textual information conveying specific features of each type of media (e.g. color for images, rhythm and tempo for music).
- Relevant keywords for each section.
- The intended effect and possibly its degree for each section/subsection (e.g. anxious, sad, optimistic...).

Possible actions might include the following action, which reduces the discussion section if this section gets longer by the equivalent time:

Duration_Action: if self:Duration>10 Then Self:Action=Fire_reduce(discussion,Duration, self:Duration10)

4.5 Design

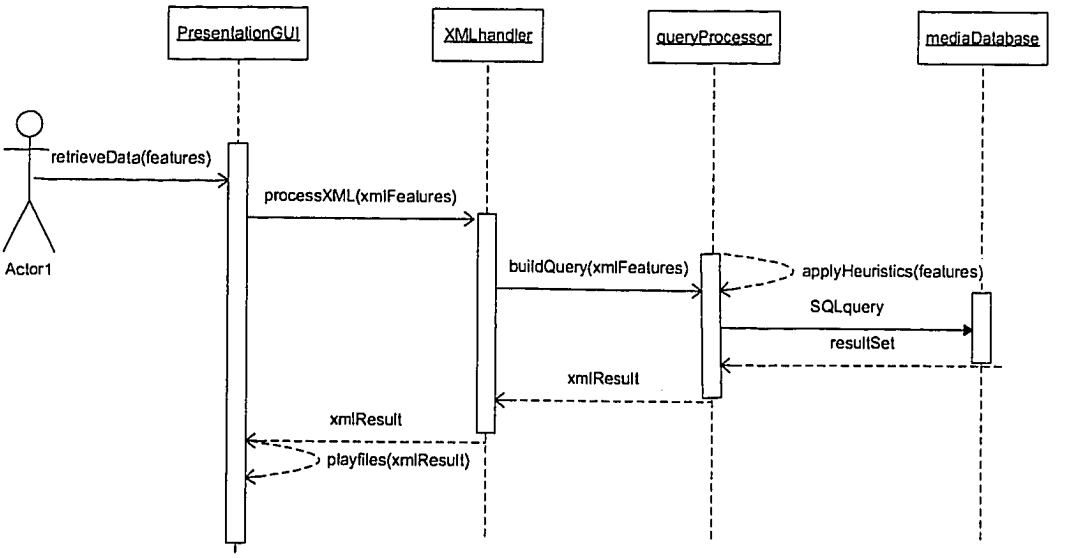

Figure 14: Sequence Diagram for Retrieve Data Use Case

Figure 14 illustrates the interaction sequence of the *Retrieve Data* use case, with the following scenario:

1. The user indicates through the presentation GUI the features of the data to be retrieved.

2. The presentation GUI reproduces the user's request in XML format and sends a message to the XML handler to process the request. An example of the XML

Figure 15: Example of XML Request

```
- <sound>
- <theme>
  <value>pk</value>
  </theme>
- <emotion>
  <value>thoughtful</value>
  <value>calm</value>
  <value>neutral</value>
  <value>intense</value>
  </emotion>
- <mdm>
  <value>music</value>
  <value>voice</value>
  </mdm>
- <mntl_spc>
  <value>poetic</value>
  <value>literal</value>
  </mntl_spc>
  </sound>
- <content>
- <clr>
  <value>red</value>
  </clr>
- <theme>
  <value>pk</value>
  </theme>
- <emotion>
  <value>thoughtful</value>
  <value>intense</value>
  </emotion>
- <mdm>
  <value>image</value>
  <value>movie</value>
  <value>text</value>
  <value>bodytext</value>
  </mdm>
- <mntl_spc>
  <value>poetic</value>
  <value>literal</value>
  </mntl_spc>
  </content>
```

Figure 16: Example of SQL Query

```
SELECT * FROM annotation WHERE (theme LIKE '%pk%') and (clr LIKE '%red%'
or clr LIKE '%blue%') and (mdm LIKE '%image%' or mdm LIKE '%movie%' or
mdm LIKE '%text%')
```

sent by the presentation GUI to the XML handler is shown in figure 15, showing elements for both <sound> (audio) requests and <content> (visual) requests.

3. The XML handler forwards the request to the Query Processor.

4. The Query Processor applies heuristics relevant to the required features.

5. The Query Processor constructs a SQL statement according to the requested features and heuristics and runs it on the media database. An example of SQL queries produced by the Query Processor is shown in figure 4.5 produced in response to the request for image/movie/text media dealing with theme 'PK' with red and blue colors.

6. The media database returns the result set to the Query Processor.

7. The Query processor translates the result set into XML format and sends it to the XML Handler.

8. The XML Handler forwards the XML result set to the Presentation GUI. Figure 17 shows an example of the result set in XML format. 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 in different manner.

9. The presentation GUI displays the files specified in the result set.

Figure 17: Result set in XML

```
<?xml version="1.0" encoding="UTF-8"?>
<Slides>
<slide URL="cody-running-small.swf" />
<slide URL="cody-walkingtowardscam.swf" />
<slide URL="endsnear.swf" />
<slide URL="pkgreendottedbckgd.swf" />
<slide URL="pkrada.swf" />
<slide URL="plexglasssquares.swf" />
<slide URL="DSCN0119.swf" />
<slide URL="DSCN0119b.swf" />
<slide URL="hair.swf" />
<slide URL="china02.swf" />
<slide URL="china07.swf" />
<slide URL="img_1890.swf" />
<slide URL="img_1912b.swf" />
<slide URL="img_1913b.swf" />
<slide URL="dscn1845.swf" />
<slide URL="dscn1898.swf" />
<slide URL="june17-031.swf" />
<slide URL="welcomming-door.swf" />
</Slides>

<?xml version="1.0" encoding="UTF-8"?>
<Sounds>
<slide URL="ea_ideaskin_pk.mp3" />
<slide URL="ea_mediatedoutput_3voices.mp3" />
<slide URL="ea_neverchangexy_pk.mp3" />
<slide URL="ea_sorryiforgot_pk.mp3" />
<slide URL="ea_withoutwords_airport.mp3" />
<slide URL="text_cody_quiet.mp3" />
<slide URL="text_inlovewithwriting_pk.mp3" />
<slide URL="text_pk_constructing.mp3" />
<slide URL="text_pk_endinguntrue.mp3" />
<slide URL="text_pk_makingmemories.mp3" />
<slide URL="text_pk_rightwords.mp3" /
<slide URL="text_worldidea_3voices.mp3" />
<slide URL="textonly_isanybodythere_pk.mp3" />
</Sounds>
```

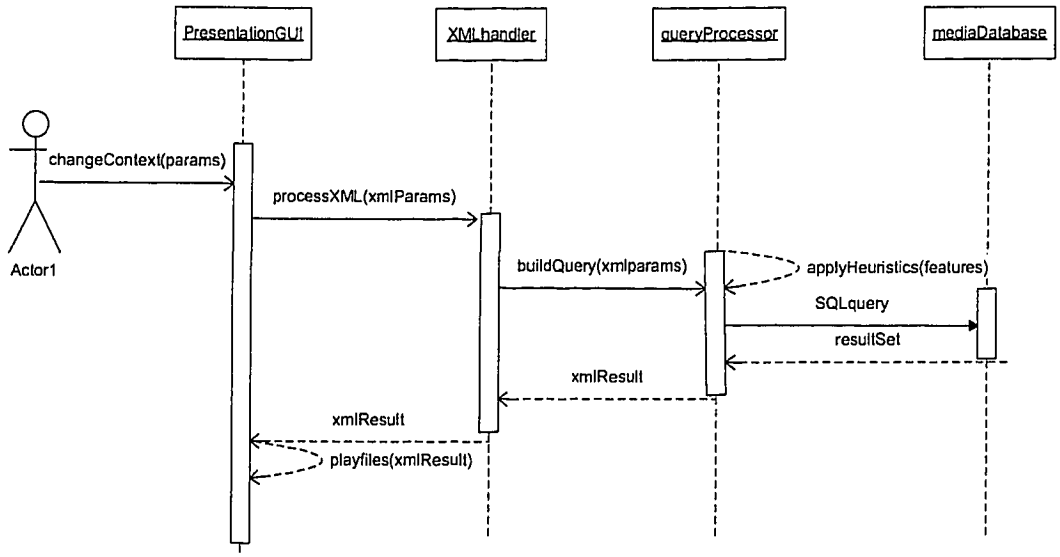

Figure 18: Sequence Diagram for Change Context Parameters

Figure 18 illustrates the *Change Context Parameters* use case, with a sequence of events very similar to the *Retrieve Data* use case. The only difference is the message sent which includes the new context parameters instead of the data features. In fact, the system handles both messages concurrently.

The *Apply Effects* use case shown in figure 19 is dealt with only through the Presentation GUI, since it is independent of the media database. It is possible however to inform the Query module of the effects, if the heuristics use this information.

4.6 The Human-Computer Interface

Depending on user requirements, the interface of a Multimedia retrieval system might offer some or all of the following functionalities: querying, browsing, continuous play, a control panel, and a relevance feedback mechanism. We add to this context specifying controls.

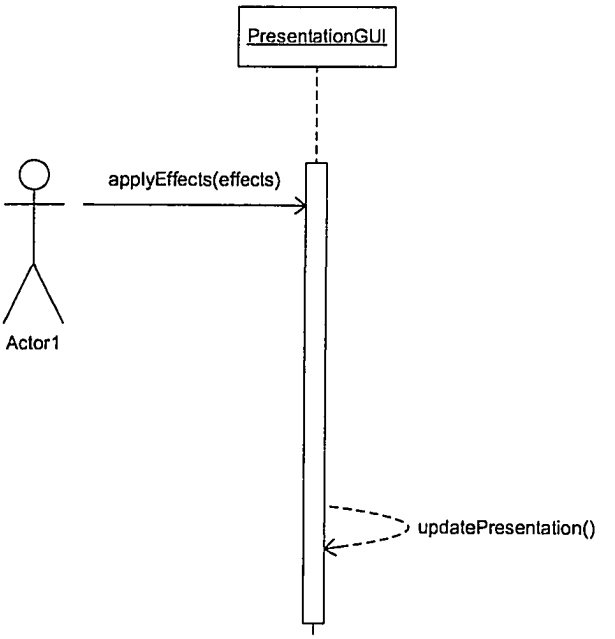

Figure 19: Sequence Diagram for Apply Effects Use Case

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.

Moreover, the goal of the interface design is to project the dialogue and the presence of the computer on stage. An interface that allows querying, triggering and browsing in real-time provides readily for enough awkwardness to be visible. Relevance feedback is a strategy that is both suitable for projecting the dialogue and enhancing the performance of the system. Visual feedback has to be stressed. The time that it takes the performer on stage to input her request or inform the system of the status and changes in the performance is part of the performance.

Notwithstanding its visibility, the user interface should be relatively fast, easy to learn and easy to manipulate. Using a set of buttons and sliders is a convenient solution for managing the interaction with the system. Each control represents one of

Figure 20: Screen-shot from the query result

the parameters of the presentation (e.g. the location in the presentation outline, the audience responsiveness, their age group, the tempo of the presentation...). Controls can also be used to capture relevance feedback from the user in the form of a rating of the selection by theme (relevant/irrelevant), medium (image/audio/video), effect (lighter/heavier), duration (shorter/longer), communicative intent (persuasive /informative)...

While a querying interface would allow the user to enter a search term, displaying the results in a visual browsing mode could provide more effective and appealing means for navigating the knowledge base. The knowledge base navigator should reflect in as much as possible its ontological hierarchies, the relationships and the links of its elements. It should adapt easily to changes in the user's focus of attention and

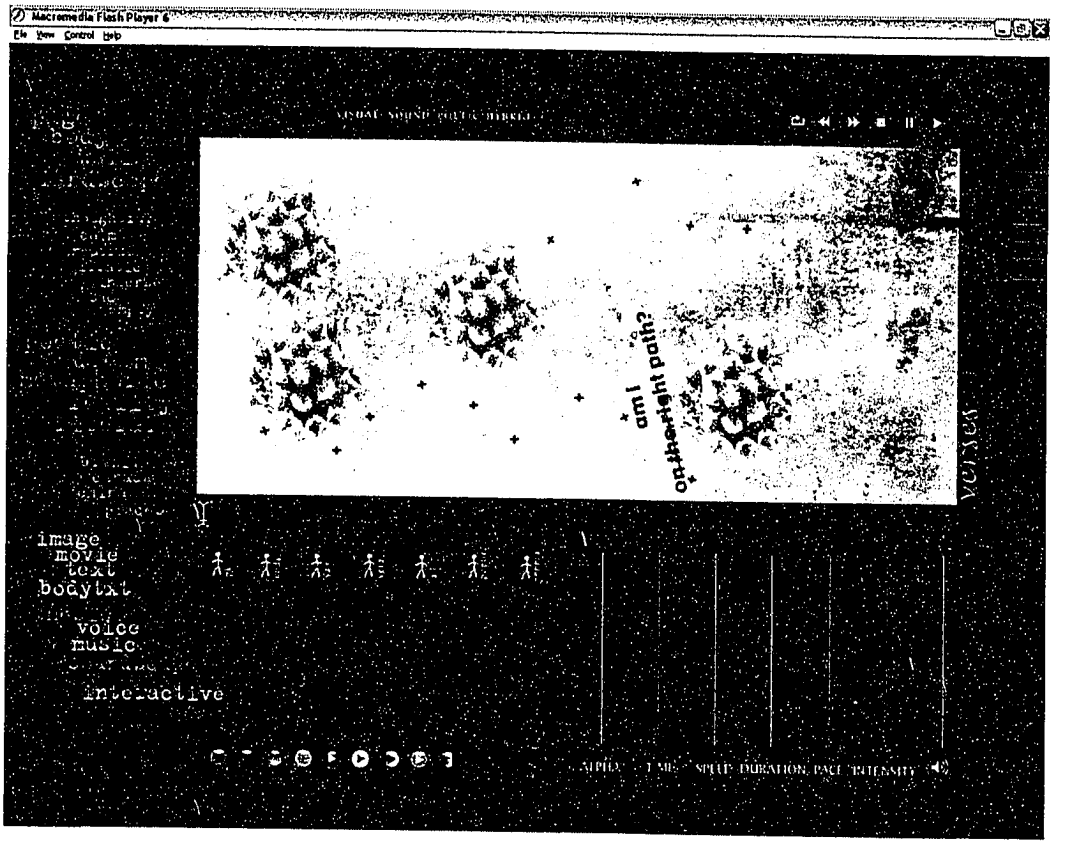

Figure 21: Screen-shot of the system's interface

browsing strategy. Displaying on demand as much information as possible about the selections without disrupting the user can be achieved through tool tip text indications.

Figure 21 shows a screen shot of the system. Through the graphical interface, the user can set any of the direct features (time, spectrum, alpha) which are either linked internally to the some features of the context model, to the data model, or apply visual effects. Using the relations (Section 3.3) and the user specifications, the most appropriate data files are retrieved from the multimedia database. From these relevant data files, only a subset may be used in the final presentation. The final selection and ordering of the data is made using the selection heuristics and the

generation patterns which make sure that the final presentation is coherent as a single production. Although not implemented yet, the framework will also allow the user to record events and playback the performance.

Chapter 5

Evaluation

Give every man thine ear, but few thy voice; take each man's censure but reserve thy judgment.
William Shakespeare, Hamlet.

Common methods for evaluating information retrieval systems focus only on measuring the effectiveness of the system [Schauble, 1997]. This criterion is further narrowed down to the relevance of retrieved documents. The influential Text Retrieval Conference (TREC) adopts this benchmark. However, according to Narasimhalu et al. [Narasimhalu *et al.*, 1998] the Information Retrieval community has been increasingly questioning this measure. In this chapter, we present first this standard approach and discuss its general limitations as well as its specific limitations in the area of multimedia and to our project. We then propose an alternative evaluation scheme based on more user-oriented measures as described by Baeza-Yates in [Baeza-Yates and Ribeiro-Neto, 1999].

5.1 Common Methods For Evaluation

Using the common approach to measuring effectiveness of retrieval in terms of the relevance of retrieved documents, the documents are divided into two mutually exclusive sets, relevant and non-relevant, as judged by a human judge. Precision and Recall figures are then calculated for the system according to the following formulae:

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

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

In practice these two figures tend to be inversely proportional: improving precision by reducing the number of documents retrieved would decrease recall and increasing recall by retrieving more documents would decrease precision. Measuring Precision at different recall levels using graphs as in Figure 22 might give a better idea about the effectiveness of the system. Some measures use a weighted formulae to favor precision or recall as in the case of the F measure calculated as follows:

$$F = (\beta^2 + 1)PR/\beta^2 P + R$$

where β =1 is neutral, gives more weight to recall for values < 1, and to precision when > 1.

Narasimhalu [Narasimhalu *et al.*, 1998] points out that Precision and Recall measures ignore such important factors as the relativity of the relevance of a document. Most Information Retrieval systems use some kind of ranking whereby documents are divided into relevant and non-relevant sets. However, this binary division of

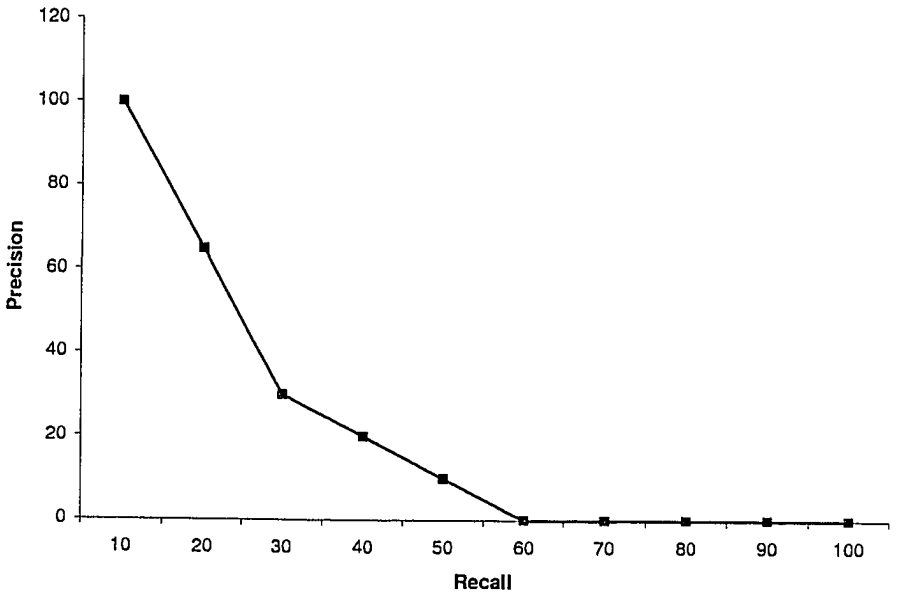

Figure 22: Precision-Recall Graph

documents is an oversimplification of the real-world. Furthermore, as Baeza-Yates [Baeza-Yates and Ribeiro-Neto, 1999] notes, Precision and Recall presume that an objective judgment of relevance or non-relevance is possible independent of the judge, an argument which can readily refuted given the discrepancies between the classification done by different judges.

5.2 Subjectivity a Necessity

The case of Multimedia Information Retrieval offers other particularities and difficulties which need to be considered for the purposes of evaluation. Objectivity is one of the hard-to-achieve goals of the evaluation of multimedia information retrieval systems. The proceedings of the TREC 2001, which included a video track for the first time, acknowledge the need for a different evaluation system for that track [Smeaton, 2001]. Narasimhalu [Narasimhalu *et al.*, 1998] indicates some of the

performance criteria not captured by Precision and Recall such as 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. He suggests that the quality of the answer is more important than the speed, and asserts the significance of ranking, order, spread and displacement in Multimedia Information Retrieval. He then proposes using fuzzy methods, neural nets or classification tree to overcome the subjectivity factor and arrive at objective measures.

Schauble [Schauble, 1997] introduces the notion of subjective relevance which hinges on the user and her information needs rather on the query formulation, since the latter might not be convenient for the task. Following this path, we can deduce that the user's participation in determining the relevance of the result is an essential factor.

5.3 Our Approach

On top of the particularities of Multimedia Information Retrieval come the application-specific ones for our project. While in general systems the material and number of relevant documents are usually unknown before hand to the user [Schauble, 1997], in this phase of the project we deal with a user with considerable prior knowledge of the material and consequently more specific expectations.

In addition, certain characteristics of our system add to the complexity of evaluation. 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. Continuous play, a requirement of the presentation is yet another factor which affects the user's perception of the system and her ability to evaluate it. At the interface level, the system has a hybrid interface which offers some querying and some browsing abilities with implicit correction through relevance feedback. All these factors change the nature of the system from a strictly information retrieval to a generative one.

We have demonstrated why the system requires a more subjective evaluation scheme. 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 are reported by Baeza-Yates [Baeza-Yates and Ribeiro-Neto, 1999] and they measure the effectiveness of the system with respect to the user's expectations. The Coverage Ratio is the portion of documents which the user was expecting to be retrieved and was actually retrieved by the system:

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

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

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

However, in order to apply these measures, the system needs to be run in slow or interrupted mode so that the user can evaluate the selections played. This potentially interferes with the information retrieved by the system, since certain heuristics are related to the speed feature. As much as possible, attention was given to making all

decisions traceable. However, the person who indexes is best positioned to evaluate — at least at a superficial level — the degree of success of the system in meeting its expectations. The direct user of the system is the performer, the annotator and their assistant(s). Indirect users are mainly the audience. If someone other than the direct users does the indexing, then this indexer's subjective decisions will impose on the system. The performer's judgment of the system, although again a subjective measure, indicates the degree to which the system meets its functional requirements.

Another method for subjective evaluation is conducting experiments with different users where they will be asked to try the system through its triggers and comparing the result with those obtained when the system is run in random mode (with no triggers activated). The purpose of these experiments is to reflect on the sense of control and relevance of returned results even for the novel user. The results of these experiments are intended to be qualitative rather than quantitative.

Finally, the audience opinion 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.

5.4 Results

Due to the Boolean retrieval model applied at this phase of the project, which only permits a relevant/non-relevant decision without the notion of relativity, the results retrieved by the system are only dependant on the annotation. We have therefore decided to postpone the evaluation to a later phase when more sophisticated IR models will be experimented with. However it is worth mentioning that a qualitative

(and subjective) favorable evaluation has been made by Prof. Langshaw, leader of the artistic group involved in the project.

Chapter 6

Conclusion and Future Work

Multimedia presentations provide a powerful tool for communication. Business, education, entertainment, tourism and culture sectors, among others, are turning to multimedia for enhanced illustration. However, in order to exploit the full potential of multimedia presentations, it is essential to allow for certain flexibility in the selection of the material for a more dynamic presentation. In a joined effort between Concordia University's Computer Science and Design Art departments, 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. In that first phase, we borrowed concepts from text analysis in order to account for the semantic dimension of the presentation. Indeed, from a semiotic perspective, multimedia presentations resemble text in what concerns their rhetorical structure and contextual analysis. We

therefore adapted the Rhetorical Structure Theory (RST) [Mann *et al.*, 1992] and the Systemic Functional (SF) linguistics model [Halliday and Hasan, 1989] for the purpose of multimedia presentations. We also proposed adopting different methods of evaluation to reflect the subjective nature of the task.

The developed system was applied to an artistic performance produced by Prof. PK Langshaw of Concordia University's Design Art Department. The outcome was deemed highly satisfactory. However, it became evident that a considerable effort would be required to adapt the system for other presentations. We therefore propose for future work developing a general architecture for multimedia run-time manipulation. This architecture would allow the replacement of the different components of the framework with relative ease. The architecture would also include tools for accomplishing common tasks such as data annotation and interface design. Research toward building this architecture will cover the areas of Natural Language Generation, Knowledge Engineering, Information Retrieval and Pattern Recognition. We expect that the outcome of the research will be a robust architecture, with an acceptable scaling and generalization capacity. Moreover, having received positive signs of interest from the museum sector, we intend to investigate the commercialization potential of the outcome and the tools developed to support the architecture.

This phase of the project started out with a task-based model, developing into a domain independent framework for adaptive multimedia presentations. The framework allows the user to define her own set of data and context features relevant to her presentation, as well as the selections heuristics and retrieval patterns. Following are some of the areas where we forsee possiblities for improvements and innovation.

6.1 The Interface

Other triggering mechanisms as speech and gesture follow naturally from the problem, providing for another layer of ambiguity. They could enhance the artistic qualities of the performance as well as provide new horizons for research.

A sophisticated cluster-based content navigation map as an alternative to triggering can be useful, especially for other related applications, like museum guides and non artistic presentations. Using this map the user would be able to navigate through the concept space and visualize the relations between the concepts.

6.2 [semi-]Automatic Annotation

Developing an annotation tool to allow the inexperienced user to do the data annotation herself will facilitate experimenting and rehearsals. This falls under the area of content-based retrieval as described in section 2.2.

6.3 Multi-User System

Another interesting possibility is a web-based multi-user version, which can be co-directed by more than a user either simultaneously in competition or in a round-robin response-triggering fashion. This would allow for some sort of collaborative production of performances in real-time. However, in order to accomplish this, several obstacles relating to the networking and media size need to be addressed first.

6.4 Using Machine Learning

Using supervised machine-learning techniques, through the employment of relevance feedback, would be a more convenient way for discovering heuristics and building user profiles. The performer would indicate during the process how relevant the selections made by the system are, thus letting the system depict the mental model of the user. Relevance feedback can also be used in real-time during the performance as a correction agent. History data collected from performances can be used as training data input to the classifiers.

6.5 The Architecture

While the architecture proposed here worked well at this phase of early experimentation with the system, it would be interesting to investigate alternative architectures like agent-based architectures.

Bibliography

[Au *et al.*, 2000] P. Au, M. Carey, S. Sewraz, Y. Guo, and S. Rger. New paradigms in information visualization. In *Proceedings of the 23rd International ACM SIGIR Conference*, Athens, Greece, 24-28 July 2000.

[Baeza-Yates and Ribeiro-Neto, 1999] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* ACM Press and Addison Wesley, 1999.

[Blum *et al.*, 1997] Thom Blum, Douglas Keislar, James Wheaton, and Erling Wold. *Intelligent Multimedia Information Retrieval*, chapter Audio Databases with Content-Based Retrieval. California:AAAI Press/ The MIT Press, 1997.

[Budzik and Hammond, 2000] J. Budzik and K. Hammond. User interactions with everyday applications as context fot just-in-time information access. In *IUI'2000, Proceedings of the 2000 International Conference on Intelligent User Interfaces*, pages 44–51. ACM, 2000.

[Carey *et al.*, 2003] M. Carey, D. Heesch, and S. Rger. Info navigator: A visualization tool for document searching and browsing. In *Proceedings of the International Conference on Distributed Multimedia Systems (DMS)*, September 2003.

[Carrer *et al.*, 1998] Marco Carrer, Leonardo Ligresti, Gulrukh Ahanger, and Thomas D.C. Little. *Multimedia Technologies and Applications for the 21st Century: Visions of World Experts*, chapter An Annotation Engine for Supporting Video Database Population, pages 161–184. Kluwer Academic Publishers, 1998.

[Champion, 2003] Erik Champion. Online exploration of mayan culture. In Hal Thwaites, editor, *Proceedings of the Ninth International Conference on Virtual Systems and Multimedia - VSMM2003*, pages 3–12, Montreal, Canada, October 2003. International Society on Virtual Systems and Multimedia.

[Chan, 2003] Chiu-Shui Chan. Virtual reality modeling of traditional chinese architecture. In Hal Thwaites, editor, *Proceedings of the Ninth International Conference on Virtual Systems and Multimedia - VSMM2003*, pages 3–12, Montreal, Canada, October 2003. International Society on Virtual Systems and Multimedia.

[Chen *et al.*, 2002] Shu-Ching Chen, Sheng-Tun Li, Mei-Ling Shyu, Chengjun Zhan, and Chengcui Zhang. A multimedia semantic model for rtsp-based multimedia presentation systems. In *Proceedings of the IEEE Fourth International Symposium on Multimedia Software Engineering (MSE2002)*, pages 124–131, Newport Beach, California, December 2002. ACM.

[Demerdash *et al.*, 2003] O. El Demerdash, PK Langshaw, and L. Kosseim. Toward the production of adaptive multimedia presentations. In Hal Thwaites, editor, *Ninth International Conference on Virtual Systems and Multimedia - Hybrid Reality: Art, Technology and the Human Factor*, pages 428–436, Montreal, October 2003. International Society on Virtual Systems and Multimedia VSMM.

[Flickner *et al.*, 1997] Myron Flickner, Harpreet Sawhney, and Wayne Nublack. *Intelligent Multimedia Information Retrieval*, chapter Query by Image and Video Content: The QBIC System. California:AAAI Press/ The MIT Press, 1997.

[Goren-Bar *et al.*, 2001] D. Goren-Bar, T.Kuflik, and T. Lavie. What do users prefer? a personalized intelligent user interface for searching information - an empirical study. In James C. Lester, editor, *IUI 2001 - Proceedings of the 2001 International Conference on Intelligent User Interfaces*, pages 65–68, Santa Fe, New Mexico, January 2001. ACM Press.

[Halliday and Hasan, 1989] M.A.K. Halliday and R. Hasan. *Context and Text: Aspects of Language in a Social Semiotic Perspective.* Oxford University Press, Oxford, 1989.

[Huang *et al.*, 2001] Lieming Huang, Thiel Ulrich, and Matthias Hemmje. Adaptively constructing the query interface for meta-search engines. In *Intelligent User Interfaces, IUI'01*, pages 97–100. ACM, 2001.

[Jaimes and Shih-Fu, 2000] Alejandro Jaimes and Shih-Fu. A conceptual framework for indexing visual information at multiple levels. In *SPIE Internet Imaging 2000*, volume 3964, pages 2–15, January 2000.

[Jelmini and Marchand-Maillet, 2003] Carlo Jelmini and Stphane Marchand-Maillet. Deva: an extensible ontology-based annotation model for visual document collections. In R. Schettini and S. Santini Eds, editors, *Proceedings of SPIE Photonics West, Electronic Imaging 2002, Internet Imaging IV*, pages 131–138, Santa Clara, CA, USA, 2003.

[Jurafsky and Martin, 2000] Daniel Jurafsky and James Martin. *Speech and Language Processing.* Prentice Hall, 2000.

[Kennedy and Mercer, 2002] Kevin Kennedy and R. Mercer. Using communicative acts to plan the cinematographic structure of animations. In Robin Cohen et al., editor, *Advances in Artificial Intelligence: Proccedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI2002, Calgary, May 27-29*, pages 133–146. Springer, Berlin, 2002.

[Layne, 1986] S. Shatford Layne. Analyzing the subject of a picture: A theoretical approach. *Cataloguing and Classification Quarterly*, 6(2), 1986.

[Leake and Scherle, 2001] David B. Leake and Ryan Scherle. Towards context-based search engine selection. In *IUI'01, Proceedings of the 2001 International Conference on Intelligent User Interfaces*, pages 109–112. ACM, 2001.

[Mann *et al.*, 1992] William Mann, Christian Matthiessen, and Sandra Thompson. *Discourse Description: Diverse Linguistic Analyses of a Fund-raising text*, chapter Rhetorical Structure Theory and Text Analysis. John Benjamins Publishing Company, Amesterdam, 1992.

[Manning and Schütze, 1999] Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, 1999.

[Marcu, 2000] Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization.* MIT Press, Cambridge, MA, 2000.

[Marti *et al.*, 1999] P. Marti, A. Rizzo, L. Petroni, and G. Tozzi M. Diligenti. Adapting the museum: a non-intrusive user modeling approach. In Judy Kay, editor,

User Modeling: Proceedings of the Seventh International Conference, UM99, pages 311–313, Banff, Canada, June 1999. Springer Wien New York.

[Maybury, 1997] Mark T. Maybury. *Intelligent Multimedia Information Retrieval*, chapter Introduction. California:AAAI Press/ The MIT Press, 1997.

[Metz, 1968] Christian Metz. *Essais sur la signification au cinéma*, volume 14 of *Collection d'esthétique*. Paris : Klincksieck, 1968.

[Narasimhalu *et al.*, 1998] A. Desai Narasimhalu, Mohan S. Kankanhalli, and Jiangkang Wu. *Multimedia Technologies and Applications for the 21st Century:Visions of World Experts*, chapter Benchmarking Multimedia Databases, pages 127–148. Kluwer Academic Publishers, 1998.

[Not and Zancanaro, 1999] E. Not and Massimo Zancanaro. Reusing information repositories for flexibly generating adaptive presentations. In *Proceedings of the IEEE International Conference on Information, Intelligence and Systems*, Washington D.C., November 1999.

[Not and Zancanaro, 2000] E. Not and M. Zancanaro. The macronode approach: Mediating between adaptive and dynamic hypermedia. In *Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, 2000.

[O'Toole, 1995] Michael O'Toole. *Discourse in Society: Systemic Functional Perspectives*, chapter A Systemic-Functional Semiotics of Art. Ablex Publishing Corporation, New Jersey, 1995.

[Panofsky, 1962] E. Panofsky. *Studies in Iconology*. Harper Row, 1962.

[Prabhakaran, 1997] B. Prabhakaran. *Multimedia Database Management Systems.* Kluwer Academic Publishers, 1997.

[Ricœur, 1969] Paul Ricœur. *Le Conflit des intérpretations - essais d'herméneutique.* L'ordre philosophique. Éditions du Seuil, Paris, 1969.

[Schauble, 1997] Peter Schauble. *Multimedia Information Retrieval:Content-Based Information Retrieval from Large Text and Audio Databases.* Kluwer Academic Publishers, 1997.

[Searle *et al.*, 1980] John R. Searle, Ferenc Kiefer, and Manfred Bierwisch, editors. *Speech Act Theory and Pragmatics*, volume 10 of *Synthese Language Library.* D. Reidel Publishing Company, 1980.

[Smeaton, 2001] A.F. Smeaton. The TREC 2001 video track report. In E.M. Voorhees and D.K. Harman, editors, *The Tenth Text Retrieval Conference, TREC 2001*, NIST Special Publication 500-250, pages 52–60. NIST, Gaithersburg, Maryland, 2001.

[Smeaton, 2002] A.F. Smeaton. The TREC 2002 video track report. In E.M. Voorhees and D.K. Harman, editors, *The Eleventh Text Retrieval Conference (Trec-11), TREC 2002*, NIST Special Publication 500-251. NIST, Gaithersburg, Maryland, 2002.

[Smith and Chang, 1997] John R. Smith and Shih-Fu Chang. *Intelligent Multimedia Information Retrieval*, chapter Querying by Color Regions Using the VisualSEEK content-Based Query System. California:AAAI Press/ The MIT Press, 1997.

[Zancanaro *et al.*, 2003] M. Zancanaro, O. Stock, and I. Alfaro. Using cinematic techniques in a multimedia museum guide. In *Proceedings of Museums and the Web 2003*, Charlotte, North Carolina, March 2003. Archives and Museum Informatics.