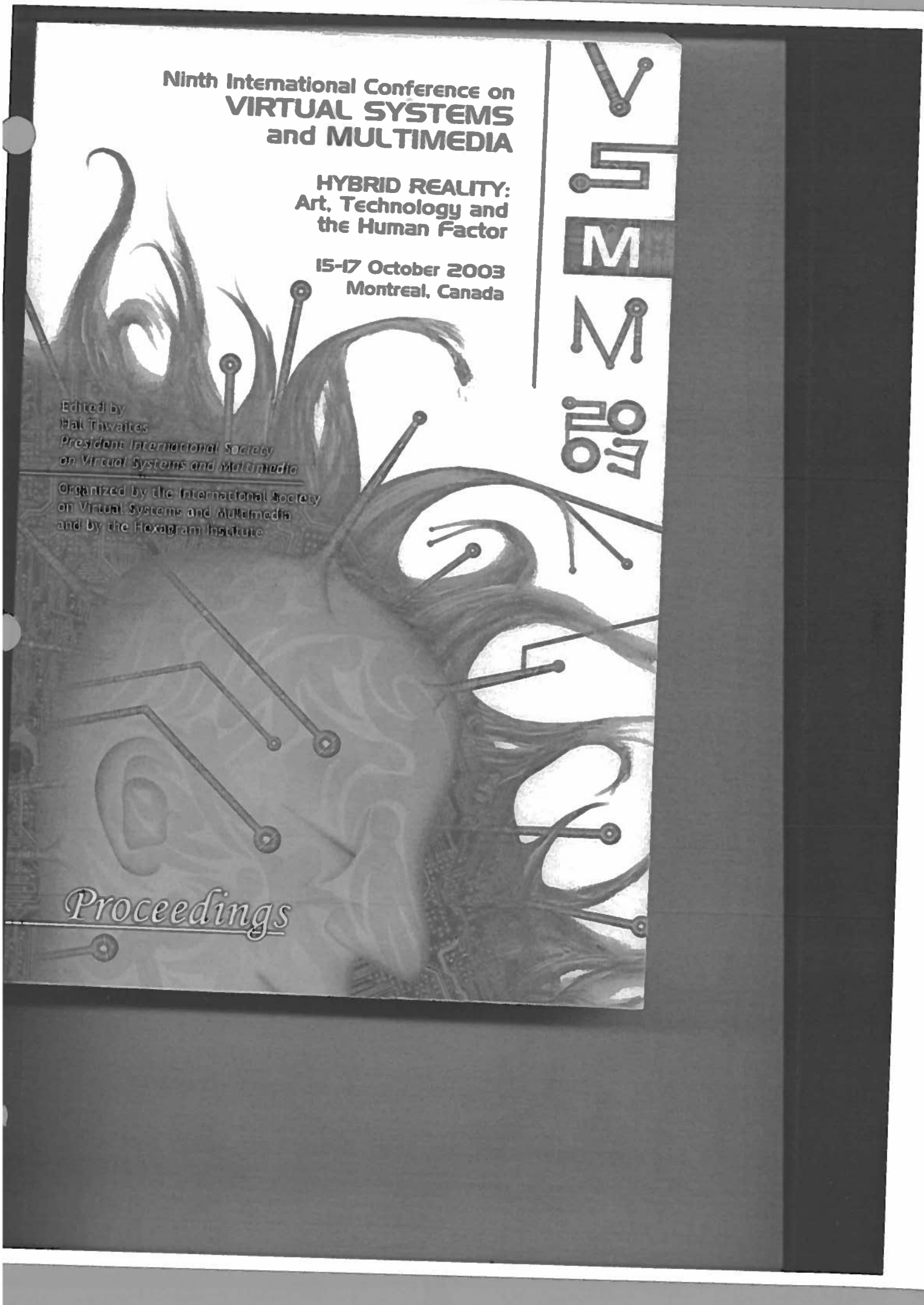
Ninth International Conference on
VIRTUAL SYSTEMS
and MULTIMEDIA
HYBRID REALITY:
Art, Technology and
the Human Factor
15-17 October 2003
Montreal, Canada
VSMM 03
Edited by
Hal Thwaites
President International Society
on Virtual Systems and Multimedia
Organized by the International Society
on Virtual Systems and Multimedia
and by the Hexagram Institute
Proceedings

Toward the Production of Adaptive Multimedia Presentations

Osama El Dermerdash*, PK Langshaw† and Leila Kosseim*
** Department of Computer Science*
† Department of Design Art
Concordia University
1455 de Maisonneuve St. West, Montreal,
Quebec, Canada H3G 1M8
osama_el@cs.concordia.ca, pkl@alcor.concordia.ca, kosseim@cs.concordia.ca

Abstract. In this paper, we propose a framework for a system that dynamically selects and plays multimedia files from a large data repository. The performance 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 performance preparation phase. We borrow concepts from the fields of discourse analysis [2] and rhetorical structure [4] as the theoretical basis for our work. To validate the framework, researchers from the Computer Science department are developing a Flash prototype with data created and annotated by a research group from the department of Design Art.

1. Introduction

Adaptive multimedia presentations involve both a preparation and a production phase. The conceptual presentation is an abstraction of what the performer has in mind as a general idea of her presentation. During the actual performance, 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 these areas as a result of the interaction with the audience or thematic pursuit of unfolding narratives.

In this paper, we propose a framework for a system that dynamically selects and plays multimedia files 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 performance preparation phase. We borrow concepts from the fields of discourse analysis [2] and rhetorical structure [4] as the theoretical basis for our work. However, in order to adapt these theories to multimedia use, as well as artistic applications, we have extended these standard models. For example, new rhetorical relations have been implemented to reflect the artistic processes, implicit and inherent in the arts domain. Also, we have recognized 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 when the end-goal is to appeal to purely scientific standards for 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.

Although we have developed our framework for the production of artistic multimedia performances, we believe that it could also be applied to produce more structured applications such as automated museum guides or adaptive instructional presentations.

2. 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 thorny and yet-to-be-resolved task of content-based retrieval (ex. [10]). However, the modelling 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.

Indeed just considering that we use more senses for interpreting multimedia data calls for a different approach for interpreting multimedia tasks. We found inspiration in Systemic Functional linguistics as described in [2]. As O'Toole illustrates through the analysis of a painting [8], the Systemic Functional model is broad enough to cover other semiotic systems, particularly visual ones.

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. In the Systemic Functional model, text is both a product and a process. Language construes context, which in turn produces language. 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.

We also draw on another linguistics theory, namely Rhetorical Structure Theory (RST) [4] for representing the possible relations between the different components of the model. RST has been used as a tool to analyse the relations between text spans of a discourse; but also as a tool to produce a coherent discourse. By selecting text spans that hold certain semantic and relations among themselves (ex. *precondition, sequence, result*…), it is possible to generate a coherent discourse from various text components. In our application, we used RST as a framework 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.

3. Artistic/Poetic Specifics

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); where interpretation still lies at a level hard to exploit.

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 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 (see [5] 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.

4. The Framework

The suggested framework for the production of multimedia presentations consists of six components: the context model, the data model, relations, selection heuristics, generation patterns and visual effects. Let us describe each component in turn.

4.1 The Context Model

In the Systemic Functional (SF) model, context refers to the environment, or any relevant features of the situation. For language, Halliday defines it as the following [2]:

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, audience, medium, rhetorical mode, mood, and history. 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. We give here a brief description of these context variables.

Outline: The outline of the presentation corresponds to the Field of Discourse in SF. It is expressed in terms of keywords. The outline can be ordered or unordered and can support timing restrictions as needed.

Time: The timeline of the presentation. Timing is a determining factor in the planning of the presentation to avoid overflows and to balance media selection.

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

Audience: The audience corresponds to tenor in SF. 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 audience require different communicative strategies, which is the case also of presenting from a position of authority as opposed to a peer-to-peer presentation.

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.

Rhetorical mode: The rhetorical mode is the strategy used at a given moment in the presentation. This is potentially dependent on the audience as well as the relative timing in the presentation. Examples given by Halliday include *persuasive*, *expository* and *didactic* [2].

Mood: 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 [3].

History: A record of selections already played is kept in history and used to avoid replaying and to balance as desired the concentration of the different media in the presentation.

4.2 The Data Model

The data model consists of the data files in any format that can be supported by Flash-MX (ex. `mp3, mpg, SWF`…) and the annotation of these with technical features and relational characteristics. As proposed by Prabhakaran [9], multimedia objects can be modelled 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 at the General, Abstract and Specific levels as applicable and the Mood of the selection. 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. name of a person, type of a chair).

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).

At this phase of the project, all annotations have been done manually. Table 1 shows an example of some data annotations according to the features explained above.

Table 1: Example of the annotation of three data files.

file name	physicality	emotion	mental space	media	Colors	keywords
DSCN0001	body (medium)	puzzled	reality	image	black/white	general:chair/snow
Text0002	landscape (big)	thoughtful	dream	text	Black	specific:Cody
DSCN0005	page (small)	puzzled	metamorphic	image	Blue	abstract:automation

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 presentation. The other purpose is to support a relational navigation map which could be used to traverse selections according to their sensory and/or semantic relations. As in the MacroNode approach [7], multiple relations can be represented. This allows for web rather than tree structures, which is customary - though not a requirement - even in the case of text structure [6]. 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 file DSCN0001.

Table 2: Example of the annotation of relation for one image file

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

4.3 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.

4.4 Selection Heuristics

The goal of the selection heuristics is to produce different interpretations of the performance 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 performer's style.

4.5 Generation Patterns

Similarly to heuristics, generation patterns are discovered while experimenting with the system during the rehearsal/preparation phase of the presentation. Patterns are complex combinations of features/heuristics. 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.

4.6 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 (fade in/out, dissolve…), scaling and zooming, etc.

5. Implementation and Evaluation

The framework described above is currently been implemented using a three-tier model. The backend operations, including applying the heuristics and querying the database are handled by a java module. The presentation module is implemented using Flash-MX and the communication between these modules is done using XML.

5.1 Implementation

Figure 1 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 4.3) 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.

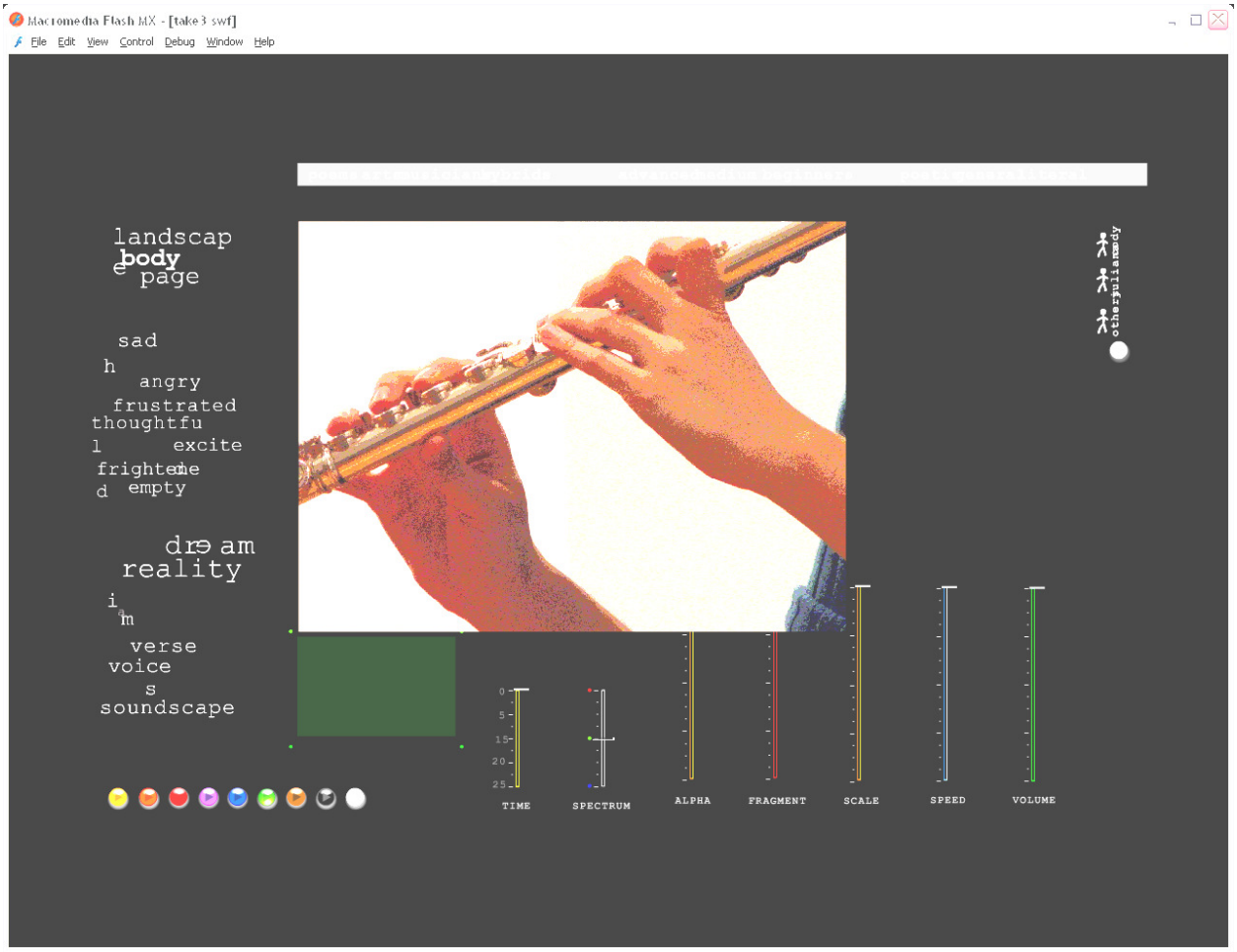

Figure 1: A screen shot of the prototype system.

5.2 Evaluation

In order to evaluate the framework, two aspects must be distinguished: the initial data retrieval which can be evaluated by scientific measures, and the performance itself, which is difficult to evaluate with scientific metrics.

The standard text retrieval systems are typically evaluated using objective measures called *precision* (the proportion of retrieved documents that are really relevant) and *recall* (the proportion of all relevant documents that the system retrieved). However, the case of multimedia information retrieval offers certain particularities and difficulties that need to be considered for the purposes 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.

Objectivity is one of the hard-to-achieve goals of the evaluation of multimedia information retrieval systems. The first two of these measures, namely the *coverage* and *novelty ratio* are reported by Baeza-Yates and Ribeiro-Neto [1]. These 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 were actually retrieved by the system, while the novelty ratio is the portion of relevant documents which were retrieved and not expected by the user. As the

prototype is currently under development, no formal evaluation has been performed yet. However, to evaluate the retrieval aspect of the system, we plan to use both the objective measures of precision and recall, and the more subjective measure of coverage and novelty ratio. These measures will be computed by running the system with a set of fixed queries and data files.

The production itself will be more difficult to evaluate since subjectivity is predominant. There does not exist a set of correct or incorrect performances to which the generated performance can be compared. The relevance and engagement of the work in artistic measure is reliant on factors such as audience reception and the performer's personal assessment.

6. Conclusion and Future Work

In this paper, we have presented a novel framework to dynamically produce multimedia presentations according to the user's preferences. The framework is grounded on theories of discourse analysis, which allowed us to consider a performance as a coherent unit rather than a series of independent multimedia data. A working prototype is currently being developed in Flash in order to demonstrate the framework's potential.

In addition to a formal evaluation of the prototype, future work includes the automatic annotation of the data, and the automatic discovery of selection heuristics and generation heuristics. So far, the annotation of the data has been performed manually. This is a very time consuming and tedious task. We believe that some more objective features can be annotated automatically by using techniques developed in pattern recognition and in text analysis. In addition, the automatic discovery of selection heuristics and generation heuristics would be a fascinating task. By using techniques in supervised learning, the system can be made to discover patterns and heuristics that are general to all users or specific to each user. In this case, a user profile can be built according to his or her artistic performing style, and loaded on demand in order to produce a customized performance to a specific artist.

This project has been a model for exploring partnership between art and science for it has striven to keep a balance of complementary relations within a hybrid and cross-disciplinary environment. In some stages, the traditional directives in the methodology, production, analysis and discourse in the two domains were contradictory and led to interesting challenges.

Acknowledgments

This work was funded by a Hexagram Institute for Research/Creation in Media Arts and Technologies and a start-up fund from the Faculty of Engineering and Computer Science. The authors would like to thank Adriana Miranda for help in data production and for designing the system interface and Dr. Sabine Bergler for many fruitful discussions on the project. Many thanks also to the students in the d_verse research group.

References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Series/Addison Wesley, New York, May 1999.

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

[3] Kevin Kennedy and R. Mercer, Using Communicative Acts to Plan the Cinematographic Structure of Animations. In: Robin Cohen *et al*. (ed.), *Advances in Artificial Intelligence: Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence*, AI 2002, Calgary, May 27-29. Springer, Berlin, 2002, pp. 133-146.

[4] William Mann *et al*., Rhetorical Structure Theory and Text Analysis. In: William Mann et al. (ed.), *Discourse Description: Diverse Linguistic Analyses of a Fund-raising text*. John Benjamins Publishing Company, Amsterdam, 1992, pp. 39-78.

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

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

[7] 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*, Trento, August 2000.

[8] Michael O'Toole, A Systemic Functional Semiotics of Art. In: Peter H. Fries et al. (ed.), *Discourse in Society: Systemic Functional Perspectives – Meaning and choice in Language: Studies for Michael Halliday. Advances in Discourse Processes*: Volume L. Ablex Publishing Corporation, New Jersey, 1995, pp. 159-179.

[9] B. Prabhakaran. *Multimedia Database Management Systems*, Kluwer Academic Publishers, Boston, MA: 1997.

[10] Alan F. Smeaton and Paul Over, The TREC-2002 Video Track Report. In *Proceedings of the Eleventh Text Retrieval Conference (TREC-11)*, NIST Special Publication: SP 500-251, Gaithersburg, Maryland, November, 2002.