Toward Formalization of Display Grammar for Interactive Media with Applications of Multimedia Information Extraction

Robin Bargar

School of Technology and Design
New York City College of Technology, City University of New York
300 Jay St., V806
Brooklyn, NY 11201
rbargar@citytech.cuny.edu

Abstract

Display grammar is a coupling of semantic and signal processing information for interactive multimedia. We propose a computational formalization binding media resource data and authoring information, including contextual data and interactive processing algorithms. We hypothesize display grammar provides a domain for mapping a range of applications of MMIE and media resource retrieval. We demonstrate a manifold model of display grammar to support this hypothesis.

Introduction

Works of interactive media indicate adoptions of new systems of production and distribution, departures from systems established for non-interactive media. Potential applications of multimedia information extraction (MMIE) for interactive media depend upon relationships between pre-recorded media resources and on-demand media signal processing. Interactive media programs are encoded as combinations of pre-produced media resources plus "*DIVA*:" *data* about the resources, the observer or the media capture context, *instructions* for computation, *variables* representing user input, and *addresses* of media resources at distributed locations. The premise of producing a single, final version of an interactive program is antiquated by user-driven on-the-fly content assembly.

MMIE applied to a media stream from an interactive media work will encounter new features and pattern combinations each time the work is observed. Given the variability of users' actions and algorithms it is desirable to develop a computational description of interactive media that combines content analysis and relevant DIVA information. To establish functional and consistent structural descriptions of interactive media we propose the analysis of the coupling of semantic structures with interactive signal processing. We introduce the concept of *display grammar* to identify this relationship.

Production Model of Interactive Media

Interactive media situate elements of program content assembly in the hands of an observer. Whereas non-interactive media content is selected and program assembly completed prior to user access. For our discussion three areas of functionality are relevant: *point-of-delivery assembly* of media programs, the *program signal* that encodes an interactive program, and the transition *from editing to authoring* as the primary process for bringing order to program content.

Interactive program signals. The term interactive media "program" is a double-entendre for a signal that carries both executable and observable information. An interactive media program signal transmits media content linked with DIVA information. The MPEG-4 standard is representative of this paradigm (Rao 2006): an end-to-end Compression-Synchronization-Delivery model of media resources transmitted as multiplexed streams of media objects encompassing both data and content, decoded under users' actions for point-of-delivery assembly within a Digital Multimedia Integration Framework (DMIF). Whereas MMIE is traditionally situated prior to the Compression phase, we will examine its potential in the Delivery phase.

Point-of-delivery assembly. Embedded computation and digital signal processing in media display devices can process data from users' actions with sufficiently low latency to provide an experience of responsive feedback. Signal processing is required for display of any digital media, including contents of media files and outputs of signal generators such as computer graphics or sound synthesis software and hardware. Signal processing is also a means of transformation of attributes of media content. We use *presentation processing* and *display synthesis* to refer to computational display of interactive media. With utilization of DSP data and processors, point of delivery may be a candidate for leveraging MMIE data that could be encoded in the program signal, and suggests MMIE inclusion in DMIF applications.

Authoring. Editing and audio mixing are manual tasks for final selection and sequencing of media resources, including effects processing such as compositing, fades and dissolves. Authoring anticipates automation of these tasks when the final presentation is executed via computation on an end-user platform; editing decisions are aug-

mented or replaced with conditional instructions that aim to ensure the quality of a variable result. To maintain production style and etiquette, editing and mixing functions may be automated on an end-user's device; thus media grammar is transformed into display grammar.

Authoring anticipates the end-user conditions of interactive reproduction including available media resources and processing capabilities. Instructions are encoded in proprietary or open source scripting and programming languages (Flash™, Powerpoint™, javascript, Dreamweaver™, MaxMSP™, and others), combining an initial selection of media resources with algorithms for further selection and ordering plus digital signal processing for dynamic display. User interaction is incorporated through programming variables that introduce data obtained from the user's device via push or pull scheduling.

MMIE representations could be provided in the authoring process to anticipate MMIE applications in a DMIF. While the MMIE might be transparent to the end user, MMIE and related pattern data could support a range of automated presentation functions that are traditionally executed manually by an editor with a keen eye and ear.

Characteristics of Display Grammar

A display grammar is a set of relationships that determine dynamic presentation processing and display synthesis, applied to but irreducible to a set of media resources. Members of a display grammar can be identified as relationships between media resources and display processes, encoded as instructions. Three characteristics are relevant for discussion of MMIE: 1. *a display grammar designates unit relationships between semantic structures and signal processing algorithms.* Included is the principle that signal processing has a semantic function when applied to media display. Relationships that make up display grammars are effectively "presentation rules" for authoring, a form of production rules. As in other grammars these production rules are structural representations of a process for rendering meaning. In a display grammar the production rules are not merely abstractions, they are also functional instructions for generating media. 2. *Computational instructions are required elements in a display grammar.* Display grammar is based upon instructions for processing and organizing media resources in connection with semantic structure. A unit in a display grammar represents one or more instructions for signal processing applied to one or more semantically-related media resources. 3. *Display grammar introduces structure that applies in common to media resources of multiple types*. In cases of combinations of media of unlike types a grammatical unit may be defined on the instructions that designate the combination of individual resources, including selection, processing and timing for presentation. A set of instructions effectively represents a complex data type comprised of multiple media resources and related dynamic processes applied. This data type is a grammatical unit of a display grammar, and maintains across multiple types of media resources irrespective of their unique and dissimilar attributes.

Semantic representation of signal processing

Formalization of display grammar requires a representation of a grammatical unit as a relation between semantic and signal processing data. Semantic and signal processing structures are inherently dissimilar in their terms of representation and implementation. A shared representation requires a construct of robust computational space and reproducible, extensible methods for incorporating dissimilar constituents in this space. We propose to minimize differences of attributes and semantic interpretation by applying a geometric representation of a signal processing space and defining correspondences to a semantic space.

Such a representation should enable media resources of unlike type to be referenced according to attributes shared in common, including semantic properties and applications of signal processing transformations. While many attributes of diverse media resources are not shared in common, a display grammar functions in a presentation space where resources of different types may be well-defined and brought into relationships under common terms of presentation such as "fade in," "cross-dissolve," "move to foreground/background," and "jump cut." Below we present a preliminary example of such a method as a demonstration of the potential functional capacity of formalized display grammars.

Characteristics of semantic structure and order. Semantic structures are discrete and may be organized in sets through relationships such as similarity, subclass/superclass or context-based association. Set relationships are not defined over quantitative order. Linear ordering of semantic terms requires the creation of a scale of values by imposing constraints on selection and interpretation of relative semantic value. A linear semantic scale may be valid within its defined context but not for semantic structure in general. Semantic structure is relevant for MMIE as a *de facto* target that guides a recognition function. Coupling ordered semantic context with signal processing dimensions could aid the extension and generalization of MMIE solutions, which tend toward optimization for narrow classes of targets.

Characteristics of signal processing structure and order. Signal processing functions are implemented as numerical expressions that constitute quantitative state spaces and can be represented as multi-dimensional systems with potentially many degrees of freedom, where each linear dimension represents a range of values for a

DSP term. This multidimensional structure lends to continuous quantitative control and ordering of signal processing functions. Coupling signal processing states with semantic sets requires interpretation of signal processing state-spaces in a particular context.

For example in a signal processing application that generates reverberation for audio signals, select combinations of parameter control values in the reverb algorithm will result in coherence of reverberation characteristics that can be described using semantic references such as "cathedral", "bathroom", or "concert hall" – terms suggestive of recognized auditory consequences of the signal processing. These examples demonstrate context and empirical observation as a rationale for linking signal processing to semantic structure; representative groupings of quantitative values in a signal processing state space take on a semantic function when applied to media resources. (Blum 1997) shows similar rationales applied in the organization of MMIE contexts for media feature target definitions.

A Manifold Representation and Interface

To investigate shared representation of semantic structure and signal processing, we apply a manifold representation of a high dimensional control space (Choi 2000). Interactive media require continuous application of signal processing, necessitating continuous control systems. A manifold representation was designed as a scalable multidimensional interface for continuous control. This tool and its underlying data representation provide a platform for demonstrating semantic coupling with signal processing control space, and examine properties of this coupling in terms of a representative functional display grammar.

Figure 1 shows a graphical user interface that represents in two dimensions a continuous and differentiable subspace of a phase space of three or more dimensions. The manifold controller (MC) applies the 2D projection of a high dimensional subspace to a 2D graphical representation with a cursor that can be positioned continuously. The MC generates high-dimensional arrays of values at each point on the control region. As the cursor moves through the control region the MC generates a continuous and differentiable series of values for each member of the array. These values are applied to real-time control of media signal processing; examples include control of sound synthesis, transformations of 3D computer graphics models, and control of multiple-axis robotic movement.

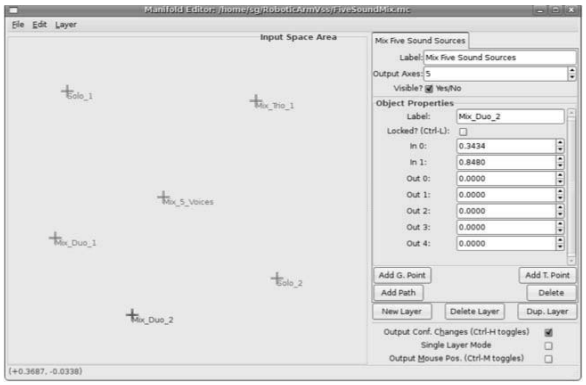

Figure 1. Generating points on a manifold control surface.

Configuring manifold control. An MC is configured by placing a set of Generating Points (GP) on a bounded plane that represents graphically a normalized 2D control region. Any GP may be positioned arbitrarily; each GP represents an ***n***-dimensional array of values associated to an ***n***-dimensional phase space. When a cursor is used to indicate a point in the control region, the MC computes a value at that point for each member of the output array, by weighting the distances of all GPs in the 2D control region relative to the position of the cursor. Altering the positions of one or more GPs will modify the resulting array values across the entire control region, with greatest effect nearest the modified GPs.

Semantic coupling to signal processing

In figure 1 each Generating Point represents an array of five floating point values normalized in the range [0, 1] and assigned to control the amplitudes of a set of five audio sound sources. One array position is designated for each member of the audio set, such that each GP represents a mix of the set of audio sources. GP "Solo_1" provides the values [1, 0, 0, 0, 0], resulting in the amplification of a single sound source and complete attenuation of the others. GP "Mix_Trio_1" provides the values [0.5, 0.5, 0, 0, 0.2], representing an audible mix of three of the sounds. Cursor positions between the GPs will produce continuous and differentiable mix levels for all sounds in the set weighted by the distance from the cursor to all GPs.

The configuration in Figure 1 represents a rudimentary coupling of semantic structure to signal processing by indexing coordinates on the control surface to designated sound sources. The coupling assumes each sound source has semantic attributes. Each GP with its associated mapping of control values to sound sources' loudness can be considered a grammatical unit in a proto-display grammar. A unit representation includes the set of sound sources, the DSP amplification function and an array of control values. Differences between units can be measured quantitatively by differences in array values and in geometric positions of corresponding GPs. However the semantic

associations are not quantifiable except by relative amplitude.

Signal processing generates semantic data. In the above example the amplitude processing of sounds provides a semantic function by generating mixes of sound sources; changes in the balance of the mix can modify the semantic references provided by the sounds. However the configuration provides limited extensibility by relying upon direct coupling of sound source and signal processing data to common geometric points and assuming the semantic attributes of the sounds. Figure 2 represents an extension of semantic functions of signal processing. A second control manifold (MC2) is defined with additional GPs that represent 9-dimensional arrays of control parameter values for reverberation signal processing applied to audio source signals. Each MC2 GP represents unique reverberation properties; as mentioned above the reverberation processing data has semantic associations to acoustic properties of recognizable spaces.

By navigating MC1 the relative amplitude of each sound source is applied to the reverberation represented by the grammatical state of MC2; separate navigation of MC2 modifies the simulated spatial qualities independently. The representation of a measurable grammatical unit can be extended to all values indexed by cursor positions on both MC1 and MC2 and their associated signal processing functions. Quantitative differences in arrays at multiple MC positions represent grammatical unit intervals. Operations upon array values generated by moving a cursor represent transformations in the grammar.

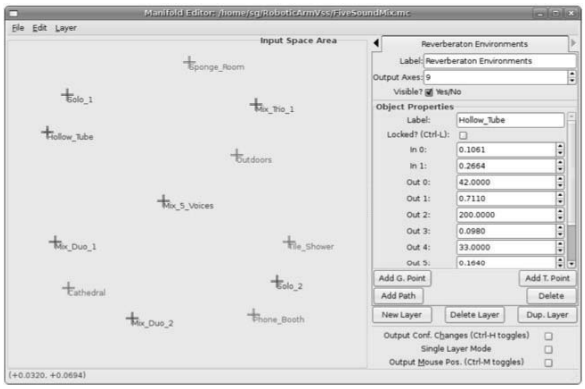

Figure 2. Two MC layers controlled by a common cursor. New GPs represent \audio reverb signal processing applied to the sound sources mixed by the MC GPs from Figure 1.

Using Ontologies to Extend Display Grammar to Media of Multiple Types

The above examples may be extended by providing additional indexing of GPs to multiple types of media such as images, video sources, and virtual camera positions in a 3D graphical scene. Media types are presented using common or related signal processing such as amplitude changes, compositing (mixing) of sources, spatial modeling and simulated movements of media sources in scenes. Arrays of control manifolds can be defined and coupled to designate grammatical structures across multiple media types and related signal processing functions.

Manual assembly and indexing of multiple manifolds and associated media sources imposes limitations in efficiency and extensibility by relying upon implicit semantic structure determined by an interface designer. Figure 3 represents an MC for interactive media that replaces preselected media sources with a semantically-based query across media of multiple types. The circles on the interface region in Figure 3 represent semantic query regions for *concepts,* which are designated search terms. In our implementation the available set of media resources are represented in an OWL file and the queries extend ontological inference across media of multiple types (Choi 2008). Overlaps of query regions are designated either as intersections or unions of concepts. GPs differentiate parameter control values for display signal processing for each type of media resource. Such a configuration contains many more dimensions than the previous examples but preserves the measurement of grammatical units across media types represented by signal processing data and common semantic terms.

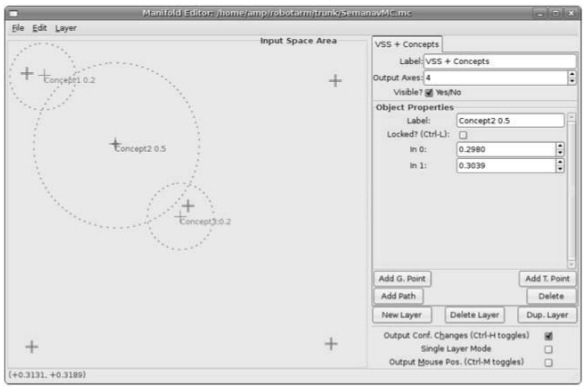

Figure 3. Semantic query regions in an MC interface.

Future Work: Developing Applications of MMIE for Display Queries and Processing

In the case represented by Figure 3, the display grammar incorporates an interactive semantic selection process and specific media resources are not known *a priori;* manifold positions designate concepts rather than fixed media resources for display signal processing. Semantic query provides a selection method to determine media resources during an interactive presentation. In such cases the grammatical unit couples display processing instructions and semantic resource selection instructions. A highly formalized display grammar can be envisioned that would

involve exclusively selection and processing algorithms without the fixed indexing of specific media resources; media resources would be incorporated downstream in the display process as a result of selection criteria provided in the grammar.

MMIE data is relevant for optimized display signal processing; media content features identified as MMIE search targets encode semantic concepts in signal processing terms. MMIE could support both semantic query and display synthesis representations in a display grammar. Semantic queries could be coupled to quantitative MMIE data representations of features as search targets. Semantic query results could be filtered according to MMIE data to select media resources that best fit a given display process, enabling MMIE-DSP criteria to contribute to the semantic selection.

Further in this vein, display processing engines could be tuned to leverage available MMIE data, for example to help determine parameter settings for individual media resources. MMIE data that identifies patterns or features could help streamline the emphasis of those features when multiple resources are synthesized into a presentation. The measure of quantitative signal information across a range of resources could enhance consistency in perceptual space such as color palette, contrast, tone, or dynamic range, by making adjustments as needed in individual resources to bring them into balance with the larger set.

At the front end of interactive media production, by relating MMIE data to display grammar the authoring of a media presentation could emphasize coherence using attribute search to support pattern-matching or complementation, rather than keyword approximations for content search and retrieval. A desirable result would be the introduction of quantifiable ranges of sensorial and semantic variety through authoring applied in the signal processing domain.

Acknowledgements

The term *display grammar* was introduced by Insook Choi, who provided helpful observations. Arthur Peters provided software engineering support.

References

Choi, I. 2000. "A Manifold Interface for Kinesthetic Notation in High-Dimensional Systems." In *Trends in Gestural Control of Music.* Battier and Wanderly, eds. Paris: IRCAM.

Choi, I. 2008. Ontologically and Graphically Assisted Media Authoring with Multiple Media Types. Submitted to *AAAI 2008 Fall Symposium on Multimedia Information Extraction*. Arlington, VA: Association for the Advancement of Artificial Intelligence.

Rao, K. R., Bojkovic, Z. and Milovanovic, D. 2006. *Introduction to Multimedia Communications.* Hoboken, N.J.: Wiley and Sons.

Blum, T., Keislar, D., Wheaton, J and Wold, E.. 1997. Audio Databases with Content-Based Retrieval. In Maybury, M. ed. *Intelligent Multimedia Information Retrieval*. AAAI Press/MIT Press. pp. 113-138.