Generic Viewer Interaction Semantics for Dynamic Virtual Video Synthesis

Craig A. Lindley

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde NSW 2113, Australia, Ph: +61-2-9325-3150, Fax: +61-2-9325-3101
Craig.Lindley@cmis.csiro.au

Anne-Marie Vercoustre

INRIA-Rocquencourt, France
anne-marie.vercoustre@inria.fr

Abstract. The FRAMES project is developing a system for video database search, content-based retrieval, and virtual video program synthesis. For dynamic synthesis applications, a video program is specified at a high level using a virtual video prescription. The prescription is a document specifying the video structure, including specifications for generating associative chains of video components. Association specifications are sent to an association engine during video synthesis. User selection of a virtual video prescription together with the default behavior of the prescription interpreter and the association engine define a tree structured search of specifications, queries, and video data components. This tree structure supports generic user interaction functions that either modify the traversal path across this tree structure, or modify the actual tree structure dynamically during video synthesis.

Introduction

The FRAMES project is developing a system for video database search, content-based retrieval, and virtual video program synthesis. The FRAMES project has been carried out within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program. Video components within the FRAMES database are described in terms of a multi-layered model of film semantics, derived from film semiotics. For dynamic video program synthesis applications, a program is specified at a high level using a virtual video prescription (Lindley and Vercoustre, 1998a). Coherent sequences of video are required, rather than just lists of material satisfying a common description. To meet this requirement, the FRAMES system uses an engine for generating associative chains of video sequences, initiated by an initial specification embedded within a virtual video prescription. Once a virtual video prescription has been selected, the prescription interpreter and associated instruction processing functions can be allowed to generate a virtual video with no further interaction from the viewer. In this case the resulting presentation has the form of a traditional linear film or video. However, depending upon the viewer's overall purpose, it may be desirable to steer the ongoing presentation in various ways. For example, the user may wish to steer the presentation towards subjects of interest and away from those of less interest, gain an overview of the area, go into detail, or follow a particular mood or emotion.

This paper defines generic user interaction semantics for dynamic virtual video synthesis based upon the data structures and sequencing functions of the FRAMES system. The semantics provide run-time interactions for the viewers of a virtual video; the interactions do not result in any permanent changes to the data structures involved, but affect the way those data structures are used to generate a particular video presentation. We begin with a summary of FRAMES system users and user tasks, provide an overview of the FRAMES system, and summarise the processes that are used to select video components during the generation of a synthesised video sequence. The high level algorithm used within the FRAMES association engine is described, and is seen to define a tree-structured search through the available video components. User interaction semantics are then analysed in terms of generic user interaction strategies, the default data structure that models the selection action of the synthesis engine, and generic interaction operations that can be defined in terms of their effect upon the implied data structure.

FRAMES System Users and User Tasks

The FRAMES video synthesis process implies four different author/system user roles that may be involved in the production and use of a virtual video. Within the FRAMES system, video data is a primitive (atomic) data input, organised as a set of discrete video sequences. The *video maker* may use a variety of software tools and products to create these digital video clips. Interactive video systems that support interaction within a complete video program represent a new medium requiring customised development of video data.

The FRAMES video synthesis engine operates upon descriptions associated with raw video data. Hence once the video data is available, a *description author* must develop a descriptor set and associate descriptors with appropriate video data sequences. The FRAMES environment includes data modeling interfaces to support this authoring process. The interfaces and underlying database are based upon the semiotic model described by Lindley and Srinivasan (1998). Once the descriptions have been created, they are stored in the FRAMES database for use by the video synthesis engine.

The FRAMES system can be used with these semantic descriptions to provide basic semantic search and retrieval services, where a user can directly interrogate the database using relational parametric queries, or interrogate the database via the FRAMES association engine either to conduct fuzzy parametric searches, or to generate an associative chain of video components. However, for many users and applications a specific high level program structure may be required. Such a structure can be defined using a virtual video prescription. A prescription, defined by a *virtual video prescription author*, contains a sequence of embedded queries for generating the low level video content, where the particular order, form, and content of the queries implements a specific type, genre and style of video production.

The final *end user/viewer* community is the audience for whom the virtual video production is created. Such a user will typically select a virtual video prescription according to their current tasks and needs, and use the FRAMES virtual video synthesis engine to generate a virtual video presentation. For dynamic virtual video synthesis, there are a number of ways and points in the process where viewer interaction is meaningful. All viewer interaction functions may be available to the authors of the interaction system, to provide feedback to authors about the appropriateness and effectiveness of descriptions and prescriptions as they are being developed. The authoring process for interactive virtual videos is highly complex, and requires careful coordination between the video makers, description authors, and prescription authors to ensure that these three levels of content are compatible and function correctly to produce coherent viewer sequences. Understanding the principles for doing this effectively is an important topic of ongoing research.

The FRAMES Video Synthesis System

The FRAMES system consists of three primary elements: a *virtual video prescription interpreter*, a *database* containing semantic descriptions of individual video components, and the *instruction engines* for generating sequences of video data. A virtual video prescription represents a high level structure of, or template for, a video program of a particular type, containing a list of instructions for generating a virtual video production (Lindley and Vercoustre, 1998a). The virtual video interpreter reads virtual video prescriptions. A user may select a prescription, which may have values assigned to various embedded parameters to reflect the particular requirements and interests of that user before being forwarded to the interpreter. The interpreter reads the instructions within a prescription sequentially, routing each instruction in turn to an appropriate processor. Three types of instructions may occur within a prescription: direct references to explicitly identified video components, parametric database queries, and specifications for generating an associative chain of video components (Lindley, 1998). *Access by direct reference* uses an explicit, hard-coded reference to a video data file plus start and end offsets of the required segment (eg. using the referencing syntax of SMIL, Hoschka 1998). *Parametric database queries* may include complex logical conditions or descriptor patterns. In parametric search, the initial query may form a hard constraint upon the material that is returned, such that all of its conditions must be satisfied. Alternatively, a *ranked* parametric search can return a list of items ranked in decreasing order of match to the initial query, down to some specified threshold. *Access by associative chaining* is a less constrained way of accessing video data, where material may be incorporated on the basis of its degree of match to an initial search specification, and then incrementally to successive component descriptions in the associative chain. Associative chaining starts with specific parameters that are progressively substituted as the chain develops. At each step of associative chaining, the video component selected for presentation at the next step is the component having descriptors that most match the association specification when parameterised using values from the descriptors attached to the video segment presented at the current step. The high-level algorithm for associative chaining is:

1. initialise the current state description according to the associative chaining specification. The current state description includes:
   - the specification of object, attribute, and entity types that will be matched in the chaining process,
   - current values for those types (including NULL values when initial values are not explicitly given or components of the next instantiation are NULL),
   - conditions and constraints upon the types and values of a condition, and
   - weights indicating the significance of particular statements in a specification
2. Generate a ranked list of video sequences matching the current state description.
3. Replace the current state description with the most highly ranked matching description: this becomes the new current state description.
4. Output the associated video sequence identification for the new current state description to the media server.
5. If further matches can be made and the termination condition (specified as a play length, number of items, or associative weight threshold) is not yet satisfied, go back to step 2.
6. End.

Since association is conducted progressively against descriptors associated with each successive video component, paths may evolve significantly away from the content descriptions that match the initial specification. This algorithm (described in detail in Lindley and Vercoustre, 1998b) has been implemented in the current FRAMES demonstrator. Specific filmic structures and forms can be generated in FRAMES by using particular description structures, association criteria and constraints. In this way the sequencing mechanisms remain generic, with emphasis shifting to the authoring of metamodels, interpretations, and specifications for the creation of specific types of dynamic virtual video productions.

Generic Interaction Strategies

User interaction in the context of dynamic virtual video synthesis can take place at several levels, and in relation to several broad types of user task. Canter et al (described in McAleese, 1989) distinguish five discernible strategies that users may use in moving through an information space:

1. scanning: covering a large area without depth
2. browsing: following a path until a goal is achieved
3. searching: striving to find an explicit goal
4. exploring: finding out the extent of the information given
5. wandering: purposeless and unstructured globetrotting

These strategies are all relevant to interaction with dynamic virtual video synthesis, and the interactive presentation system for virtual videos should support each strategy. To these five strategies we can also add:

6. viewing: allowing the algorithm to generate a video sequence without further direction from a user (ie. the viewer is passively watching a video)

Dynamic virtual video syntheses in the FRAMES project uses the viewing model as the default behavior of the system. That is, once a virtual video prescription has been selected, the synthesiser generates the video display based upon that prescription and the semantics defined by the underlying algorithms. The virtual video prescription may define a video program amounting to a scan, browse, search, exploration of, or wander through the underlying video database, depending upon the application-specific purpose of the prescription. To provide interactive viewing functions, suitable interfaces must be provided allowing viewers to modify the behavior of the video synthesis engine away from this default behavior within the form defined by the original virtual video prescription.

User Interaction Semantics

A prescription can be customised for a particular user by setting its parameter values. Parametric search may be an exact search mechanism (eg. if a traditional relational database is used), or may involve a fuzzy search process that returns identifiers of video component having descriptors that approximately match the search query, ranked in decreasing order of match to the query. A video synthesis system incorporating ranked search can include interfaces allowing users to select from the ranked list of returned results. Associative chaining can be modified in several ways by user interactions, by using user interactions to effectively modify the chaining specification dynamically as chaining proceeds. Users can modify the entity types used to associate the current component with the next component, modify the current entity values, set or reset constraints upon entity values, or modify the weightings upon entity types. Users can also interrupt the default selection of the most highly associated video component by selecting another ranked element as the current element, which will re-parameterise the associative chaining specification at the current point in the chain.

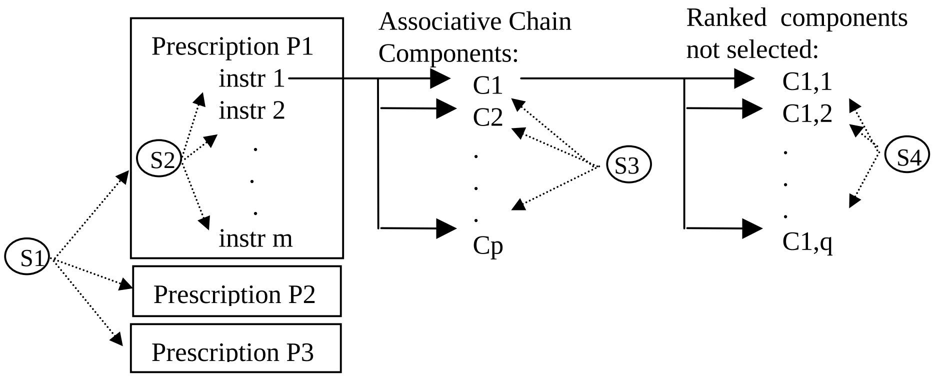

Fig. 1.

The semantics of these user interactions can be modeled by regarding the operation of the association engine as a tree search behaviour, as shown on Figure 1. In considering choices that can be made by users, it is useful to regard the starting point as the decision about which virtual prescription to execute, this being the root node of the search tree. Each prescription contains a list of instructions that constitute its child nodes. The algorithm that interprets prescriptions will execute each instructions in sequential order. An instruction (specifically, an instruction that is an association specification) generates a series of video components that are its child nodes in turn, each component being presented for display in the sequential order in which it is returned. Finally, for each selected video component in a series, there is a list of other candidate components that have not been selected, ranked in decreasing order of associative strength (to the *previous* component in the selected list); this ranked list may be considered to be a set of child nodes for a selected component. Hence the video synthesis process defines an ordered, depth-first traversal of the system data structures and the dynamically generated association structure of video components.

The default behavior of the synthesis engine without user interaction satisfies the user interaction strategy identified above as *viewing*. However, to support scanning, browsing, searching, exploring, and wandering strategies, specific and generic interaction functions can be provided. These are divided into two classes. The first class of interaction functions are those that determine the path taken by the user in traversing the default synthesis tree amount to functions that interrupt or modify the default depth-first traversal behavior of the algorithm. These functions include:

- control of whether the process should stop, loop back to some point (eg. as identified on a history list), or proceed to the next default item

- jump to a position on the tree other than the next position defined by the depth-first algorithm
- display a set of video components in parallel

The second class of interaction functions are those that dynamically alter the structure of the default tree during video synthesis are functions that effectively produce an alteration in the specification that is driving the generation of a virtual video production. This can include:

- functions that dynamically modify virtual video prescriptions (eg. changing the values of variables used within a prescription during execution)
- functions that dynamically modify queries prior to their execution, or as they are being executed. Examples include adding or removing descriptor types that associative matching is taking place against, and modifying the weightings attached to descriptor types.

Related Work

Interactive video covers a broad range of technologies and interests, including interactive video editing systems, model-based video image generation, and interactive search and browsing of video data in archives or databases. The FRAMES project is addressing the interactive use of predefined video sequences. Dynamic access to predefined video using content-based retrieval techniques has generally been based upon an information retrieval model in which data is generated in response to a single query (eg. the IBM QBIC system, http:// wwwqbic.almaden.ibm.com/ stage/ index.html); sequencing from this perspective is a contextual task within which content-based retrieval may take place. The MOVI project has incorporated some automated video analysis techniques into an interactive video environment that then uses hard-coded links between video elements (see http:// www.inrialpes.fr/ movi/ Demos/ DemoPascal/ videoclic.html). Unlike these approaches, FRAMES generates links between video sequences dynamically using an associative chaining approach similar to that of the Automatist storytelling system developed at MIT (Davenport and Murtaugh, 1995, and Murtaugh, 1996). The Automatist system uses simple keyword descriptors specified by authors and associated with relatively self-contained video segments. In Automatist, users can interact with the associative chaining process either by explicitly modifying the influence of specific keyword descriptors arranged around the periphery of the interface, or by selecting a less strongly associated video component to become the current displayed component determining the ongoing associative chain. The FRAMES system extends this associative chaining approach by using a highly structured semantic model (described in Lindley and Srinivasan, 1998), which allows greater discrimination on descriptor types, and more types of relationship between sequenced video components. Flexible and modifiable association specifications in FRAMES and the incorporation of direct references and parametric queries in high level prescriptions create opportunities for interaction beyond the simple selection of keywords and ranked components.

Conclusion

This paper has presented an analysis of the underlying semantics of user interaction in the context of the FRAMES dynamic virtual video sequence synthesis algorithms. Ongoing research is addressing the presentation of interaction options to users, and the problem of disorientation within the unfolding interactive video.

References

Aigrain P., Zhang H., and Petkovic D. 1996 “Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review”, *Multimedia Tools and Applications* 3, 179-202, Klewer Academic Publishers, The Netherlands.

Davenport G. and Murtaugh M. 1995 “ConText: Towards the Evolving Documentary” Proceedings, ACM Multimedia, San Francisco, California, Nov. 5-11.

Hoschka P.(ed) 1998, “Synchronised Multimedia Integration Language (SMIL) 1.0 Specification” W3C Recommendation 15 June 1998.

Lindley C. A. 1998 “The FRAMES Processing Model for the Synthesis of Dynamic Virtual Video Sequences”, Second International Workshop on Query Processing in Multimedia Information Systems (QPMIDS) August 26-27th 1998 in conjunction with 9th International Conference DEXA98 Vienna, Austria.

Lindley C. A. and Srinivasan U. 1998 “Query Semantics for Content-Based Retrieval of Video Data: An Empirical Investigation”, Storage and Retrieval Issues in Image- and Multimedia Databases, August 24-28, in conjunction with 9th International Conference DEXA98 Vienna, Austria.

Lindley C. A. & Vercoustre A. M. 1998a “Intelligent Video Synthesis Using Virtual Video Prescriptions”, Proceedings, International Conference on Computational Intelligence and Multimedia Applications, Churchill, Victoria, 9-11 Feb., 661-666.

Lindley C. A. & Vercoustre A. M. 1998b “A Specification Language for Dynamic Virtual Video Sequence Generation”, International Symposium on Audio, Video, Image Processing and Intelligent Applications, 17-21 August, Baden-Baden, Germany.

McAleese R. 1989 “Navigation and Browsing in Hypertext” in *Hypertext theory into practice*, R. McAleese ed., Ablex Publishing Corp., 6-44.

Murtaugh M. 1996 *The Automatist Storytelling System*, Masters Thesis, MIT Media Lab, Massachusetts Institute of Technology.