Generic Film Forms for Dynamic Virtual Video Synthesis

Craig A. Lindley
CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde NSW 2113, Australia
Craig.Lindley@cmis.csiro.au

Abstract

The FRAMES project within the RDN CRC (Cooperative Research Centre for Research Data Networks) is developing an experimental environment for video content-based retrieval and dynamic virtual video synthesis from archives of video data. The FRAMES research prototype is a video synthesis system incorporating an association engine for automatically generating associative sequences of video data. Authors preparing video descriptions, association specifications, and component video sequences require guidelines for developing the primitive video components and semantic representations required to create particular film forms. High level formal models of film and video provide such a methodology. In particular, categorical, associational, and abstract film forms can be generated based upon video component descriptors and specifications, while narrative and rhetorical forms require the introduction of independent rules or relations expressing causal and rhetorical relationships, respectively.

1. Introduction

The FRAMES project within the ACSYS CRC (Advanced Computational Systems Cooperative Research Centre) is developing an experimental environment for video content-based retrieval [7] and dynamic virtual video synthesis [4] from archives of video data. The FRAMES research prototype is a video synthesis system incorporating a multi-level database model for describing the semantics of archived video data, together with an association engine for automatically generating associative sequences of video data. This paper addressed the methodological question of which kinds of video semantic descriptors and specifications should be used to generate particular forms of synthetic (or virtual) video. The paper begins by characterising the video synthesis problem and describing the FRAMES research demonstrator. We then present a set of high level models of film form. A detailed example is developed to demonstrate how these formal models are related to the FRAMES multi-level model of video semantics, and to show how virtual videos having those forms can be generated using the FRAMES association language and chaining engine. While some film forms can be created in a direct way using the current FRAMES prototype, other forms require supplementation of the basic semantic content models with general models of causal and rhetorical relationships, providing requirements specifications for extensions to the initial FRAMES association engine and semantics database.

2. FRAMES Virtual Video Synthesis

Approaches to video synthesis can be divided into two broad categories. The first category includes approaches that seek to synthesise all of the image content (eg. the Stanford Virtual Theatre Project). The second category includes approaches that seek to synthesise coherent video sequences from pre-existing sequential components. The FRAMES system is a research prototype for exploring video semantics representation and dynamic virtual video synthesis from pre-existing sequential components. In the FRAMES system, a *virtual video prescription* provides high level content, structuring and pacing control. A virtual video prescription is essentially a virtual video program template containing a series of instructions that include association specifications for initiating and controlling the generation of associative sequences of video data [4]. An association specification is expressed in a propositional language that includes initial values for video content descriptors (ie. semantic annotations), constraints upon descriptor values, and weights upon descriptor types [5]. Association specifications can be modified dynamically for active control of pacing and progression of detail, and also as a mechanism by which a viewer can interact with the dynamic virtual video synthesis process.

Video semantic descriptors are modelled in FRAMES using a metamodel based upon the film semiotics pioneered by the film theorist Christian Metz [6]. This novel metamodel includes five levels of cinematic codification [7]. The *perceptual level* is the level at which

0-7695-0253-9/99 $10.00 © 1999 IEEE

visual phenomena become perceptually meaningful, and visual distinctions are perceived by a viewer. The *diegetic level* is the level at which the basic perceptual features of an image are organised into the four-dimensional spatio-temporal world represented by a video image or sequence of video images. The *cinematic level* is the level of formal film and video techniques incorporated in the production of expressive artefacts ("a film", or "a video"). This level includes camera operations (pan, tilt, zoom), lighting schemes, and optical effects. The *connotative level* of video semantics includes metaphorical, analogical, and associative meanings that the denoted (ie. diegetic) objects and events of a video may have. Finally, the *subtextual level* of video semantics includes more specialised meanings of symbols and signifiers. Modelling "the meaning" of a video, shot, or sequence requires the description of the video object at any or all of the levels described above. The different levels interact, so that, for example, particular cinematic devices can be used to create different connotations or subtextual meanings while dealing with similar diegetic material.

Virtual videos are produced in FRAMES based upon the interpretation by the virtual video engine of a virtual video prescription created by an author [4]. An instruction within a prescription can provide access to video components via direct reference, parametric database search, or by associative chaining. *Associative chaining* is a method of automatically selecting and sequencing video data on the basis of its degree of match to an initial search specification and then incrementally to successive component descriptions in the associative chain. Since association is conducted progressively against descriptors associated with each successive video component, paths may follow semantic chains that progressively deviate from the initial matching criteria. Specific filmic structures and forms can be generated by the FRAMES association engine based upon a suitable association specification [5]. 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.

3. High Level Syntactic Forms for Video

Alternate approaches to video synthesis raise the question of which models and algorithms are appropriate to use for particular interactive or dynamic virtual video productions. In particular, the question arises of what can be accomplished using only atemporal descriptors associated with separate video components, and when is it necessary to introduce temporal representations or representations about meanings created only by the combination of separate components. Davenport and Murtaugh [2] emphasise narrative as the primary organising principle for evolving documentary. However, narrative is only one among a number of organising forms of filmic material. Bordwell and Thompson [1] identify the additional categories of *categorical, associational, abstract*, and *rhetorical* forms. Each of these forms represents a different (partially codified) syntactic structure for film sequences. In most real films, the forms apply at multiple levels of film structure, a given film sequence may involve multiple forms at the same level, and multiple forms may occur at different levels. In this section we describe each form, and illustrate (where appropriate) how it might be generated from FRAMES semantic annotations and the FRAMES associative chaining algorithm using the example database shown in Table 1. The database contains video sequences having the content descriptors, classified by level of codification, shown on the table.

Table 1. Example Database.

Shot Number	Diegetic	Perceptual	Connotative
1	- mountain - cloud	- blue	- tranquil - skiing - wolves
2	- mountain - rain	- grey - blue	- gloomy - tranquil
3	- foothills - rain	- green	- gloomy - tranquil
4	- foothills - river	- brown - blue	
5	- town - rain	- grey	- gloomy
6	- town - flood	- brown	- disaster

Descriptors used in the example are typed as follows:

-diegetic descriptors are limited to *Object* and *Location* types
-*Location* types include: mountain, foothills, and town
-*Object* types include: cloud, rain, river, and flood.
-perceptual descriptors are limited to *Colour* types
-connotative descriptors are generically typed

In general descriptors are not absolute, and are authored together with knowledge of the database contents, process instructions, virtual video specifications, and possibly rules to create various forms of virtual video. Connotative and subtextual descriptors may vary widely for a particular shot if provided by different authors. The annotations created at these levels may reflect a strong authorial intent within a particular virtual video production - they do not necessarily represent "what the author thinks" about a particular shot or its content, but how the author wants that shot to function within a specific production.

3.1. Categorical Form

Categorical films use subjects or categories as a basis for their organisation, typically basing each segment of the film on one category or subcategory. Common examples of stereotyped categorical films include lifestyle and gardening programs, travelogues, and sporting programs.

The FRAMES data model associates typed annotations with video segments. Annotation types are category types, and the annotations themselves are category names. An associative chain is initiated by sending an association specification to the association engine. The specification includes the descriptor types to chain on, as well as initial values and possible constraints upon values. For a categorical film, the supercategory or general topic (if specified) can be represented by a constrained (and hence unchanging) descriptor value. The subcategories to move through are then represented as unconstrained descriptor types. The rate at which the categories change can be determined by a weighting attached to the subcategory descriptors: the higher the positive weighting, the more slowly the categories will change, while the more negative the weighting, the faster the categories will change.

As a demonstration of this, a virtual video can be generated from the example database described on Table 1. If the topic of interest is rainfall, and we wish to move quickly through the different locations as subcategories, the specification "(object == "rain" and location[-0.5] = "mountain") may generate the sequence of clips numbered: 2, 3, 5, showing rain on the mountain, in the foothills, and in the town, respectively. The first clip will be clip 2, since it is the only clip that matches both the constrained value of object == "rain" and the initial value of location = "mountain". After the first clip, the constrained object value persists as "rain", but the location value will be altered. With this very simple specification, either of the clips 3 or 5 could come next, and the overall order of clips after the first clip will depend upon ad hoc aspects of the implementation, such as the de facto order in which records are retrieved by the database management system. This is acceptable in the case of a categorical film, since we are only concerned that the categories of interest are selected, and not in their order or selection.

3.2. Associational Form

Associational videos suggest expressive qualities and concepts by grouping images. The juxtaposition of images in an associational film leads the viewer to look for some connection between them, an association that binds them together. This form is distinguished from rhetorical and narrative forms, in that the binding association is not a causal or rhetorical relationship. Repetition of motifs is particularly important in associational forms.

Within the FRAMES system, an associative relationship between video sequences amounts to having a common descriptor value. As an example, we may wish to generate a sequence of clips manifesting the concept of gloom. This can be done using a very simple specification of the form: "(connotation == "gloom")", which will generate a sequence containing the clips 2, 3, and 5, showing mountain rain, rain in the foothills, and a flooded town, again in an unspecified order.

Note that there is a strong involvement of this form of associational relationship within films of the categorical form: a categorical film has a high level structure determined by representing a number of desired categories, while *within* each category there may be a number of separate clips associated by membership of the common category. Moreover, the set of categories itself may be associated as subcategories of a single supercategory, as in the previous example where all of the clips are associated with the concept of rain.

3.3. Abstract Form

In films having an *abstract* form, the audience's attention is drawn to abstract visual and sonic qualities of the things depicted (shape, colour, rhythm, etc.). In other words, films based upon the abstract form are films in which the essential subject is the film medium itself, and the phenomenological effects that the medium can produce other than by simply representing things in the world (or arguments, etc.). Objects represented in abstract films are selected for their visual qualities and not for their worldly functions as the objects represented by the image. Many experimental or avante garde films are concerned with the exploration of the abstract or formal properties of the film medium. Abstract qualities are also taken into account in other film forms, but do not usually provide the dominant principles of their organisation.

The specification "(colour == "blue")" is an example of an associative specification involving abstract descriptor. The execution of this specification by the FRAMES assocaition engine, using a database of video annotations that includes colour descriptions, will create a sequence containing all of the blue shots in the database, ie. sequences 1, 2 and 4. Alternatively, the specification "(colour[-0.5]=blue)" will create a sequence starting with a blue shot, and then containing a series of shots that vary rapidly in colour from one shot to the next.

Abstract films can use either persistent (positively weighted) or rapidly changing (negatively weighted) descriptor values. Hence they can actually be either categorical or associational films, as defined above and in relation to the specific formal aspects represented in a category or associative feature at the cinematic or physical level of codification. Their distinguishing characteristic as abstract films is in the use of non-representational structuring principles. The abstract form is therefore an orthogonal classification which may be distinguished from more representational functions of categorical and associational films.

3.4. Narrative Form

Narrative and rhetorical film forms are both distinguished from categorical, associational, and abstract films by their creation of meanings that specifically require the sequential association of initially distinct video sequences. That is, any basic video component in a categorical or associational film will represent its designated meaning (ie. associated descriptor) irrespectively of what precedes or follows it. For rhetorical and narrative films, however, the rhetorical and narrative meanings created by the sequential juxtaposition of basic video components is not, and generally cannot, be conveyed by those components in isolation (unless the components manifest representational redundancy).

Different theorists focus on different qualities that characterise narrative, such as the conjunction of events and recounting (narrative is regarded as a kind of knowledge), teleology (and hence a sense of closure) and wholeness, or temporality. Basically though, narrative is about telling a story, and hence involves a system of causally interrelated events, actions, and situations. Commercial dramatic films are narrative films, although narrative organisation also appears frequently in many forms of documentary.

Narrative is concerned with the creation of a pattern of cause-effect relationships among the diegetic events, actions, and situations depicted by a film. In general this pattern may be extremely complex (eg. involving converging and diverging patterns of implied parallel action), can involve the words and motivations of characters, and may be narrated using time orders quite different from those depicted in the diegetic world (eg. using flashbacks or flashforwards). As Davis [3] notes, the representation of action and causality can draw upon a large body of research in reasoning about action, and an understanding is needed of the level of resolution required in the decomposition of represented action. Here a very simple method is considered for synthesising coherent narrative video sequences. Even in this very simple example, representations are required beyond the semantic annotations needed to create synthetic videos having categorical, associational and abstract forms. This is because the description associated with each separate component does not represent the unique meanings (ie. implications of causality) created by the combination of those components.

For the example database described above, a narrative sequence can be created using a simple forward-chaining inference engine and the following two causal relations (or rules):

(location = mountain and object = rain) causes
(location = foothills and object = river)
(location = foothills and object = river) causes
(location = town and object = flood)

The chaining algorithm begins with the video component showing rain in the mountains. With each inferential step the algorithm matches the latest effect with the cause of another effect. The components associated with each effect state are selected in the same order that they occur in the cause-effect sequence. In this case this results in the sequence of clips 2, 4, and 6. This sequence shows: rain in the mountains followed by water flowing in the river in the foothills, then the town being flooded. The implied narrative is very clear, that the rain in the mountains created the water that ended up flooding the town.

The creation of such narrative sequences can be reasonably straightforward if the contents of the database are known, and for a comparatively constrained range of narrative possibilities. For a less constrained database, the creation of coherent narratives quickly becomes problematic. Narrative generation incurs general problems of reasoning about action, such as the frame problem: how do we know the full extent of what should or should not follow from the performance of a given action in a given context. Filmic narratives raise additional problems, such as how to decide how much of the video data identified as the causal consequences of a starting clip to actually display, ie. what are the relevant consequences and what are irrelevant in a given situation? These questions are topics of ongoing research.

3.5. Rhetorical Form

Rhetorical films present an argument and lay out evidence to support it. The aim of such a film is to persuade the audience to hold a particular opinion or belief. Rhetorical films will frequently present arguments as if they are observations or facts, and will typically fail to present any opposing views. A standard description of rhetorical form suggests that it begins with an introduction of the situation, goes on to a discussion of the relevant facts, then presents proofs that a given solution fits those facts, and ends with an epilogue that summarises what has gone before. Common examples of rhetorical films are television commercials.

The representations required to create simple rhetorical sequences are analogous to those required for narrative sequences. For narrative, diegetic state descriptions are associated causally with other diegetic state descriptions. For rhetorical films, diegetic, connotative, or subtextual state descriptions are associated by various rhetorical relationships to other diegetic, connotative, or subtextual state descriptions. Here we consider simple syllogistic implication. In particular, the argument may be expressed informally that if it rains in the mountains, then the river will rise and the town will flood. In terms of the annotations available in the example database, this argument could be represented by the rules:

If (location = mountain and object = rain)
then (location = foothills and object = river).

If (location = foothills and object = river)
then (location = town and object = flood).

An inference engine could begin with the consequence (location = town and object = flood) and backward chain through (location = foothills and object = river) to the conditions (location = mountain and object = rain). Video components can be found that satisfy each of these conditions, and the sequence presented in either the forward or reverse order to the direction of inferential search. In the example database, reversing the search order will result in the sequence of clips 2, 4, and 6, the same as that derived in the narrative case, but this time using logical implication rules.

4. Conclusion

The identification of the five film forms discussed in this paper clarifies the kind of virtual video production that can be achieved using the existing FRAMES matching process, and provides functional requirements for extensions of the basic matching process to include both causal and rhetorical relationships. Further research is addressing the detailed development of causal and rhetorical representations, techniques for incorporating causal and rhetorical information into the matching process, and techniques for the systematic creation of nested formal structures.

5. Acknowledgements

The author wishes to acknowledge that this work was carried out within the Cooperative Research Centre for Research Data Networks established under the Australian Government's Cooperative Research Centre (CRC) Program and acknowledge the support of the Advanced Computational Systems CRC under which the work described in this paper is administered.

References

[1] D. Bordwell and K. Thompson, *Film Art: An Introduction*, 5th edn., McGraw-Hill, 1997.

[2] G. Davenport and M. Murtaugh, "ConText: Towards the Evolving Documentary" Proceedings, ACM Multimedia, San Francisco, California, 1995, 5-11.

[3] M. Davis, "Knowledge Representation for Video", *Proceedings of the 12th National Conference on Artificial Intelligence*, AAAI, MIT Press, 1994, pp. 120-127.

[4] C. A. Lindley and A.-M. Vercoustre, "Intelligent Video Synthesis Using Virtual Video Prescriptions", Proceedings, International Conference on Computational Intelligence and Multimedia Applications, Churchill, Victoria, 1998, pp. 661-666.

[5] C. A. Lindley and A.-M. Vercoustre, "A Specification Language for Dynamic Virtual Video Sequence Generation", International Symposium on Audio, Video, Image Processing and Intelligent Applications, Baden-Baden, Germany, 1998.

[6] C. Metz, *Film Language: A Semiotics of the Cinema*, trans. by M. Taylor, The University of Chicago Press, 1974.

[7] U. Srinivasan, C. Lindley, and B. Simpson-Young, "A Multi-model framework for Video Information Systems", accepted for "Semantic Issues in Multimedia Systems", 8th IFIP 2.6 Working Conference on Database Semantics (DS-8), Rotorua, New Zealand, 1999.