AIIDE 2012 Doctoral Consortium
AAAI Technical Report WS-12-18

Toward a Narrative Comprehension Model of Cinematic Generation for 3D Virtual Environments

Bradley A. Cassell
Liquid Narrative Group
North Carolina State University
Raleigh, NC 27695
bacassel@ncsu.edu

Abstract

Most systems for generating cinematic shot sequences for virtual environments focus on the low-level problems of camera placement. While this approach will create a sequence of camera shots which film individual events in a virtual environment, it does not account for the high-level effects shot sequences have on viewer inferences. There are systems which are based on well known cinematography principles such as the rule of thirds and other framing principals, however these usually utilize schemas or predefined shots and do not reason about the high level cognitive effects on the viewer. In this paper a system is proposed which can reason directly about these high-level cognitive and narrative effects of a shot sequence on the viewer's mental state.

Research Problem

Human cinematographers plan shot sequences either explicitly or intuitively to manipulate the mental state of the viewer (Branigan 1992). The narrative processes and techniques used to design these shot sequences are similar to those used to construct written discourse. Research also has shown that people use the same mental processes to understand both written and cinematic discourse (Magliano, Miller, and Zwaan 2001). One way to increase the expressiveness of cinematic generators is for them to reason about these cognitive and narrative processes used by the viewers. Specifically, I want to focus on how to automatically generate cinematics which can non-verbally communicate the internal mental state of a character when he or she is deliberating on changing his or her plan of action using models of narrative understanding. A key element to cinematics like these is that they include camera shots which seem redundant, or don't present new information about the environment to the viewer. For a system which focuses solely on low-level constraint solving for placement, it is difficult to generate shot sequences which contain shots that don't present new or important information. These shots can be key to foregrounding information important to the unraveling story. Cognitive psychology research has shown that people can build situation models while they are reading or watching a narrative, and that the events in the models are connected along various indecies (Zwaan and Radvansky 1998; Magliano, Miller, and Zwaan 2001). Using these situation models allows for measuring how salient, or easily recalled, prior events are compared to the current event in a story. My research will develop methods for shot generation that will reason explicitly about these situation model dynamics in order to effectively communicate these changes in character mental state. Richer and more expresses cinematics will be able to be generated with a system using these methods which I posit will lead to better viewer experiences.

Proposed Research

Most work in cinematic generation reasons about low-level frame-by-frame placement of a camera in a virtual scene (Bares et al. 2000; Christianson et al. 1996; Tomlinson, Blumberg, and Nain 2000). However, in film, cinematographers either explicitly or intuitively build shot sequences to manipulate the mental state of the viewer (Branigan 1992). Initial work on addressing this planning process for automated cinematic generation within 3D virtual environments was done by Jhala and Young (Jhala and Young 2010). Their approach draws on methods like those of Moore and Paris (Moore and Paris 1993), Mann and Thompson (Mann and Thompson 1988) and others from the natural language generation community where hierarchical templates are used as an approach to discourse structure generation. Similarly, the goal of my research is to develop a system which can generating cinematic sequences for stories within virtual environments from a discourse perspective using the natural language methods in Darshak. Specifically, the system will generate shot sequences that can effectively communicate the mental states of a character when he or she is deciding to change his or her course of action within a story. In particular I consider cases where characters' observations of their world and internal reflection cause them to change their plans.

An important type of shot used in the situations that I am focusing my attention on is a shot that seem redundant. Redundant shots are shots which show events (flashbacks) or properties of objects in the story world (a door being open for example) which have already been shown in previous shots. How these shots can be used is discussed in later sections.

In this paper I present arguments for the importance of

explicit models of both story and discourse goals in the generation of effective narrative, and I also give a description of preliminary work that employs these models in an automatic cinematography system that conveys intention dynamics of story characters.

Previous Approaches

To date, there have been two primary types of automatic cinematic generation research: low-level control of the camera to place it and film objects correctly, and high-level reasoning about narrative pragmatics. Most work in automatic cinematic generation has focused on the low-level control of the camera. These approaches mainly use pre-determined cinematic constraints as guides to shot sequence creation (Christianson et al. 1996; Bares et al. 2000). Separate from this cinematic work, systems for textual generation have been developed that could be useful in cinematic generation (Mann and Thompson 1988; Moore and Paris 1993; Walker 1993). However, little work has been done on designing a system which generates shot sequences based on their pragmatic effect on viewer inferencing.

Christianson et. al. (1996) describe a system which uses cinematic idioms – standard sequences of shots that communicate specific actions – to generate a cinematic sequence. The idioms are formalized in the Declarative Camera Control Language developed by Christianson et. al. The system uses these idioms to construct shot sequences based on cinematography principles. However, it still relies on these idioms instead of reasoning about the desired effect on the viewer directly. Bares et. al. (2000) have developed a system which allows the user to specify particular camera shots by creating storyboards. The system then takes the storyboards and uses a constraint solver to determine where to place the camera. This, however, still leaves the high-level shot composition up to the author of the storyboards. The cinematography system created by Tomlinson et. al. (2000) is a reactive system rather than a plan-based one. It makes use of an intelligent camera agent that films characters based on how much the characters want to be filmed. Each character will communicate how important it believes it is, and the camera agent will use this to help determine when to switch between the characters and what type of shot to use. Cambot (Elson and Riedl 2007) is an offline system that reads in cinematic constraints from a script, and then generates cinematic sequences based on the information in the script. The script includes information such as characters, actions, temporal specifications, and location, blocking, view and scene constraints. In addition to the script, Cambot makes use of handwritten cinematic domain knowledge that includes information on stages, blockings, and shsot types. Cambot searches through all valid blocking and stage combinations and finds the best location on the set. It then compiles the shots that are needed to cover each beat, or single action, in the script into reels and ranks the reels. the final cinematic is the reel with the highest ranking.

Computational linguists and others working in the area of natural language generation have developed systems that generate mult-utterance text. The system presented by Moore and Pairs (1993) takes the approach that a text generation process needs to reason about the intentions of the utterer as well as the rhetorical structures of the text to effectively communicate the content of the text, especially when a system is participating in a dialog. Their system can build text hierarchically based on satisfying intentionality goals and appropriate rhetorical structure.

Walker (1993) developed a system that could generate text that contained informationally redundant utterances (or IRU's). These utterances' informational content has already been presented, but repeating it has an effect on the inference of the text by the reader or hearer. Her system accomplishes this by incorporating a working memory model of the hearer, which was adapted from prior work by Landauer (1975). This memory model allows for the information stored in it about the story world to fall out of context; IRU's serve to bring it back into focus at the appropriate time.

Cinematic generation systems that use text generation methodologies have been successful in producing cinematics based on high level narrative pragmatics. The Darshak system (Jhala and Young 2010), for instance uses hierarchical planning to generate cinematics using abstract shots that are composed of sequences of more primitive shots. These abstract shots are shots which can be viewed as narrative principals such as *Deliverance*. This particular abstract shot has three participants, an unfortunate character, a threatening character, and a rescuing character. Primitive shots are those which are used to actually film these three characters, a close up for example. While Darshak does guarantee a limited model of salience based on composition of the abstract shots, it does not have mechanics to reason directly about what prior events outside of the decompositions for the abstract shots might be salient or not.

Current Work

In my work I plan to create a system that can generate shot sequences that effectively communicate character mental states during story events, but which reasons directly about the cognitive effect on the viewer. To do this, I plan on incorporating narrative theories and cognitive models of narrative comprehension with a hierarchical discourse planner. A system such as this will need several basic parts:

1. Representations for story and camera plans
2. Model of the viewer's mental state
3. A planner that can use this mental state model to generate shots sequences that effectively manipulate the same mental state model

First, the system will need to have representation for both the story and the discourse. Plan representations have been shown to be very useful in both areas (Jhala and Young 2010; Riedl and Young 2010; Young 1999). The plan structures used for the story and the cinematic discourse will be similar to each other and connected causally after the discourse planner has run. The plan structure for the story will have to change somewhat from traditional planning. Ways to represent and maintain character mental states at particular events in the story will need to be added. Initial work

will focus on the discourse generation and will only take the story plan as input, not generate it. Future work will focus on how to generate the required story plans that contain the events needed for the discourse generator. However, currently I am planning on using a simplified belief manager to model character beliefs during the story generation process. This would allow for maintaining character beliefs and for reasoning about them during the planning process. In addition, I present the idea of a *deliberation* action. This is a story action that only has preconditions and effects involving the character beliefs. These deliberation actions would model a character's ability to consider their current plan of action and possibly change it. Coupled with the belief manager these deliberation actions would give the planner a way to manipulate character beliefs to construct a story that contains character deliberation.

Next, the viewer's mental state at any given time during the discourse plan will need to be modeled and reasoned about. This will allow the discourse planner to form shot sequences which intentionally manipulate the viewer's mental state to communicate narrative goals. Cognitive models have been developed that characterize both how people store narrative information and how people navigate through their mental state. Event-Indexing situation models (EISM) are one such cognitive model (Zwaan and Radvansky 1998). With situation models people are said to form a mental model of the narrative as they read it. The model is updated as more information is presented to the reader. The model and the events contained in it are associated and updated along 5 indicies: time, space, causality, intentionality and protagonist. These indicies define how salient events in the model are to one-another and how easy it is for the viewer to mentally navigate from one to the other within the situation model.

Another cognitive model is the attention/working memory model presented by Walker (1993) which was adapted from Landauer (1975). This model stores information about the world as beliefs of the reader and allows for these beliefs to fall out of focus by the use of a memory decay. As the beliefs fall out of focus they move further and further away from the foreground of the memory model, effectively making them more difficult to recall. Beliefs are also able to be brought back into focus by foregrounding them again. This is achieved by the use of what Walker calls *Redundant Utterances*. These are utterances that do not present new information about the world, yet have specific uses such as foregrounding beliefs or ideas between participants in a dialog. I extend this idea to redundant shots, which are camera shots which add no new information to a viewer's mental model, yet are important in that they can foreground previous information. Using redundant shots, relevant events or outcomes of events at any earlier part of a story that are not salient for the viewer could be made salient for the viewer, thus foregrounding knowledge that is important to the unfolding action.

A computational model of the event-indexing situation model has been proposed in previous work (Cardona-Rivera et al. 2012). This computational model extends the plan structure of intentional partial order causal link (IPOCL) plan structures (Riedl and Young 2010). The plan structure is augmented to contain data for all five indices of the EISM. Plan steps are required to contain entries for location, time, and protagonist. Most steps have an entry for location, however in this an entry is forced. Time is treated as time periods thus allowing multiple steps to be connected in the same time index. This computational model also requires that the IPOCL plan be totally ordered, but only after generation. The protagonist entry is a flag indicating if one of the steps variables is binded to the protagonist. IPOCL plan structures already contain information on causal and intentional relationships between plan steps. The story plans which are input to my system will be in this EISM augmented IPOCL representation. This gives the initial situation model of the viewer for the planner to reason about.

For a planner to use situation models it will need to maintain and manipulate a predicted situation model that the viewer has while watching the cinematic. The base algorithm to form the sequence of shots will be a decompositional partial order causal link planner to make use of the hierarchical structures Darshak uses. The goal of the planner would be to have the viewer's situation model of the story in the correct configuration at the end of the shot sequence. Each abstract Darshak action in the plan would be a shot or shot sequence that modifies the viewer's situation model along one or more of the 5 event-indicies from situation model theory. Each shot action would add or remove support for information held within the situation model and move it closer to the goal situation model. As these camera actions get added to the plan the planner will find story actions in the story plan input to attach them to that contain the events in the camera action similar to what Darshak does. If the story plan contains the deliberation actions mentioned earlier, the planner can check to see how salient the events supporting the knowledge needed for this deliberation action are in the viewer's situation model. If they are not salient, the cinematic generator can make sure that the actions supporting the character's belief preconditions for that deliberation action are salient in the viewer's situation model. The end state of the plan would be when the situation model contains specific information which was the communicative goal of the cinematic. Thus, cinematics which choose shots to bring specific ideas to attention can be created.

Planned Work

An important research issue is how to use the planner to construct the shot sequence. Traditional planning typically works from the goal state backwards, moving through plan space instead of world space. This method poses a problem when using a mental model. Ideally the mental model is manipulated as planning occurs, however there is no knowledge of previous mental state if the planner starts from he goal state. One solution to this would be to treat the mental state as additional planning constraints, however this moves away from the natural progression a mental state would have while watching a narrative from beginning to end. Another solution would be to use a forward chaining planner. This would require an accurate heuristic to guide the planner through plan space.

Several other problems still remain to be solved with this approach. First, specific weights for each index in the EISM are needed as well as a threshold for salience. This, however, is a limitation of the EISM; the discourse planning mechanics prsented here could function with any system that accurately determines salience of prior story events. Another issue is exactly how to foreground prior actions. In this context planning actions will be shots and foregrounding could be accomplished by showing certain preconditions of the shot actions. This would bring into focus the events and objects necessary for the current shot action to be understood causally. However, what it means to show a precondition needs to be determined. This could be as simple as a close up of an object that is the predicate of the precondition, however it may also require a full flashback. Also, not all preconditions require explicit communication to the viewer. If all of the preconditions are shown it could result in confusion, since some may not be extremely relevant. Preventing knowledge is another challenge. Some cinematics, specifically those of the mystery genre, may require withholding certain information. For example, when filming a murder scene, the cinematographer must select shots that effectively convey the action in the scene without identifying the murder's face and giving away his or her identity. This would require restricting working memory so that it does not contain this information.

Targeted cinematic sequences for this system to produce are sequences that contain specific instances of shots explaining character deliberation. One example would be when a character observes a specific property of an object in the story world that makes them discontinue their current course of action because it has some special meaning to them. A gift from another character that makes the main character realize they should stop harmful actions directed at that character for example. The goal with this type of situation would be to have camera shots be generated that show the property of the object the character is observing and explain why it is relevant to them, thus explaining their choice to discontinue their course of action.

Conclusion

Current cinematic generation systems are able to produce effective cinematics, but they lack the ability to reason directly about the effects the cinematics have on a viewer's mental state. I have presented support for the use of cognitive and natural language techniques for cinematic generation. A cinematic generator which can reason about these elements of narrative understanding will be able to generate custom cinematics which achieve specific narrative communicative goals. At first, my research will focus on how to construct cinematic sequences which can explain a character's mental model at the time of deliberation of action, however I will later extend this to be a general model of the construction of shot sequences that communicate a character's internal mental state in any situation.

References

Bares, W.; McDermott, S.; Boudreaux, C.; and Thainimit, S. 2000. Virtual 3D Camera Composition from Frame Constraints. *Proceedings of the Eighth ACM International Conference on Multimedia* 177–186.

Branigan, E. 1992. *Narrative Comprehension and Film*. Routledge.

Cardona-Rivera, R. E.; Cassell, B. A.; Ware, S. G.; and Young, R. M. 2012. Indexter: A computational model of the event-indexing situation model for characterizing narratives. In *In the working notes of the Workshop on Computational Models of Narrative at the Language Resources and Evaluation Conference*, 32 – 41.

Christianson, D.; Anderson, S.; He, L.; Salesin, D.; Weld, D.; and Cohen, M. 1996. Declarative Camera Control for Automatic Cinematography. In *Proceedings of the National Conference on Artificial Intelligence*, 148–155.

Elson, D. K., and Riedl, M. O. 2007. A Lightweight Intelligent Virtual Cinematography System for Machinima Production. In *Artificial Intelligence and Interactive Digital Entertainment 2007*.

Jhala, A., and Young, R. M. 2010. Cinematic Visual Discourse: Representation, Generation, and Evaluation. *IEEE Transactions on Computational Intelligence and AI in Games* 2(2):69–81.

Landauer, T. 1975. Memory Without Organization: Properties of a Model With Random Storage and Undirected Retrieval. *Cognitive Psychology* 7(4):495–531.

Magliano, J. P.; Miller, J.; and Zwaan, R. a. 2001. Indexing space and time in film understanding. *Applied Cognitive Psychology* 15(5):533–545.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3):243–281.

Moore, J., and Paris, C. 1993. Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. *Computational Linguistics* 19(4):651–694.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 164–7.

Tomlinson, B.; Blumberg, B.; and Nain, D. 2000. Expressive Autonomous Cinematography for Interactive Virtual Environments. In *Proceedings of the Fourth International Conference on Autonomous Agents*, 317–324. ACM.

Walker, M. 1993. *Informational Redundancy and Resource Bounds in Dialogue*. Ph.D. Dissertation, University of Pennsylvania, The Institute for Research in Cognitive Science.

Young, R. M. 1999. Notes on the use of plan structures in the creation of interactive plot. In *AAAI Fall Symposium on Narrative Intelligence*, 164–167.

Zwaan, R. A., and Radvansky, G. A. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123(2):162–85.