Some Assembly Required : Making a "Smart" Editor

by Andrew Rogers Beechum

6.199 : Advanced Undergraduate Project

Supervised by : Professor Glorianna Davenport

Media Lab, Interactive Cinema

May 24st, 1996

Table of Contents

Abstract

Film is a powerful visual medium that transmits an artists concept of a narrative through the use of images and sounds arranged according to complex cinematic languages. *Some Assembly Required* (SAR) is a system that attempts to solve the problem of dumb cropping that occurs in the film to video transfer process today. Utilizing the methodology of structured video, SAR allows a director to control a dynamic editing process via annotations and the rules of Master Scene Language. This dynamic editing preserves the vital visual information that the director deems important to his narrative.

We have produced a structured video movie, *Two Viewpoints* to demonstrate the usability of SAR.

My evaluation of the current implementation of SAR is that the resulting product, while answering some questions, still leaves room for further development, specifically in the area of narrative intelligence.

SAR was developed in conjunction with David Tamés. A more detailed discussion of the theory behind SAR can be found in Tamés' Master's Thesis, "Some Assembly Required: Cinematic Knowledge-Based Reconstruction of Structured Video Sequences."

1. Introduction

Friday night, the tie is off, your feet are up, and your evening has been catered by Dominoe's and for your entertainment pleasure a Blockbuster Video. You play the movie and as the FBI warning glides by there's a brief disclaimer that most people ignore, "This movie has been reformatted to fit your screen." A typical weekend night, but this message signals the destruction of hours of effort and the mangling of the very visuals that comprise film.

Film is a stream of carefully constructed images and sounds that combine to provide the viewer with a vital and enthralling story. Once movie screening was the special province of movie theaters and movie production the province of Hollywood studios, but the home theater revolution with its VCRs and inexpensive cameras has increased the public's knowledge and awareness of film as a communication medium. At the same time fundamental differences in the display mechanisms for film and video have produced a problem of conversion. The current methods of converting from film to video are either inadequate or destructive. Meanwhile increased computer technology has spawned new ideas for construction and display of film and video.

This paper examines aspects of the ***Some Assembly Required*** **(SAR)** tool kit, a system that explores the intelligent re-editing of film and video sequences. SAR is designed to reorganize a sequence in a manner that communicates the important visual and narrative information to the viewer. To illustrate the concepts of SAR, we have also filmed and built a structured video sequence, *Two Viewpoints.*

First there is a brief overview of background concepts, such as film to video transfering, structured video, master scene language, and *Moviemaker's Workspace.* I will then explain the SAR tool kit and focus on the **Sequence Orchestrator** , the tool responsible for outputting a reworked sequence.

2. Background

2.1 Film to Video Transferring

On the surface film and video are matching media streams. They both consist of dominant video tracks with a concurrent audio track. However, viewed side by side the difference becomes plainly visible. The two frames below illustrate the difference in **aspect ratio** of film and video. Aspect ratio is defined as the relationship between the width and the height of the image. Television, on the left, is standardized as a 1.33:1 ratio. (i.e., a TV screen 1 inch high would be 1.33 inches wide) Unlike television film has multiple aspect ratio standards. Films produced before the 1950's typically were shot in a 1.33:1 ratio, but with the development of wide screen technology a myriad of ratios appeared. Over time 1.85:1 has become the U.S. standard while 1.66:1 has become Europe's standard.

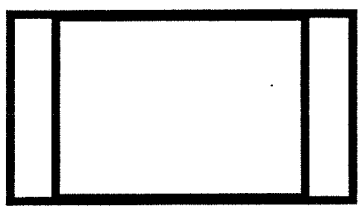

Figure 1. 1.85 vs. 1.33

The effect of the wider screens for film produces large problems during the video transfer. What do we do with the image in all of the excess screen space? The most commonly used methods for tackling this question are **letterboxing** and **pan and scan cropping**.

Letterboxing reduces the amount of screen height used for the image by placing black bars at the top and bottom of the screen. This method eliminates enough of the height to produce a 1.85:1 ratio. (see Figure 2) However, by reducing the usable screen real estate images may be squeezed too much in the remaining screen space and be unrecognizable. Hollywood's reluctance to release letterboxed movies except as collector edition productions effectively renders letterboxing useless.

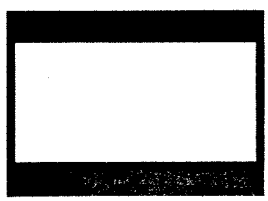

Figure 2. Letterboxing

Pan and scan is the most commonly used method for cropping film to fit video. Essentially, an editor chooses portions of the screen that are 1.33:1 and crops away the rest of the frame. This has the advantage of keeping details understandable and prominent in the frame, but suffers because a large amount of the frame is thrown away. Today on film sets you will hear directors and cinematographers talking about "shooting for video." What this means is the film is framed in a 1.33:1 aspect ratio. The rest of the 1.85:1 frame is extraneous space. This begs the question of why even bother shooting for 1.85:1? 1.85:1 is still shot for a variety of reason including economics and artistic.

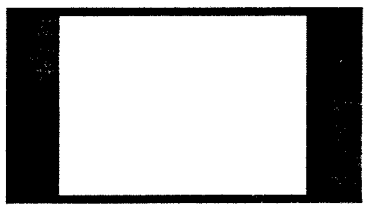

Figure 3. Pan and Scan Cropping

2.2 Structured Video and Cheops

As currently viewed, video is a 2D stream of flat pixels transmitted at 30 frames per second. Once a video frame is created it is difficult to change. These frames lock us into a particular presentation and incur a high storage cost. Structured Video is a concept being explored at the MIT Media Lab by V. Michael Bove and others. In Structured Video the cinematic frame is built dynamically from a collection of objects. We would have set objects, actor objects, sound objects among others. Structured Video presumes that these objects and their behavior is transmitted to your future television set which will use digital processing power to composite them into a single image. Structured Video moves away from static, clunky frames to smaller, more manipulable objects.

Structured Video allows us to step outside of the world of aspect ratios, framings, and frame rates. We now have a mixture of 3 dimensional elements (a set) and 2 dimensional elements (actor objects) that we can dynamically control and composite. Framing, in a Structured Video environment, is no longer static. Framing can be controlled dynamically by a processor inside your television. If a character doesn't quite fit in an frame, the character or the aspects of a camera can be altered to fix the framing problem, or in parlance, "maintain the

shot." At the Media Lab, researchers have built a special machine, Cheops, that performs the compositing required in a structured video environment.

SAR models a structured video environment with digital compositing and editing. SAR also overlays on the compositing process a notion of a shot based language. This shot based language allows directors to specify images in traditional Master Scene Language definations.

2.3 Master Scene Language

As with any technically complex occupation, film has a special language associated with it. The language most commonly associated with the slick Hollywood film is the Master Scene Cinema Language (MSL) which D.W. Griffith developed. MSL provides a standard method of talking about film, both framing and editing. The framing language used to discuss film provides a clear verbal communication of a visual idea. As an example a director can tell a cinematographer to shoot a "medium shot of Rick" and the cinematographer has a clear idea of the director's intention. MSL also consists of a standard way of shooting and editing a scene that provides the film maker's needed coverage. The primary method of coverage is to produce a Master shot, a wide shot that captures the entire scene in one frame, and subsequent shots which capture progressively smaller framings to provide detailed examination of the film space. This coverage insures greater flexibility in post production when the film is being edited together.

2.4 Moviemaker's Workspace

Moviemaker's Workspace is a previsualization tool conceived and developed by Scott Higgins. Using a three dimensional graphics environment containing information about a set and video images of actors, Moviemaker's Workspace allowed a director to develop a storyboard of a film sequence by combining 3D set models and 2D views of the actors. Camera placement and movement was specified through control panels that communicated choices via standard Master Scene Language terms as described above. This environment combined with a simple playback mechanism allowed the director to explore different shot compositions and sequence edits without a huge expenditure in

film. The full technical and theoretical background is discussed in Higgin's Master Thesis.

The playback mechanisms that I developed for theMoviemaker's Workspace were the basis for the playback mechanism described in section 4.

3. SAR Overview

3.1 The SAR Toolkit

Some Assembly Required (SAR) is a tool kit designed to aid in the creation, annotation, and playback of structured video sequences. The project consists of several stages and tools. There is a production stage, where certain camera setups are maintained for the creation of the Structured Video movie. There is a suite of post-production tools, aiding the construction and manipulation of a Structured Video environment. Finally SAR itself is composed of two tools, a Shot Composer, and the Sequence Orchestrator. The latter is discussed in Section 4.

3.11 Production - *Two Viewpoints* and beyond

In order to prove that Structured Video productions could benefit from an added cinematic language, we needed to work with a structured video sequence to re-edit on the fly. With the help of Tony Romain and the Cheops group we produced a structured video version of Romain's adaptation of Charlotte Gilman Perkins' *Yellow Wallpaper.*

The script was rewritten to simplify the story and accentuate the differences between the two main characters, Kathy and John. The differences between the characters allowed us to divide the story into two stories, one from Kathy's perspective, the other from John's.

In order to simplify the production of our actor object we video taped our actors against a chromakey blue screen in a studio. In this production we used five synchronized Betacams. Each of these cameras were time synchronized so that a time of 00:00:00:00 (0 hours, 0 minutes, 0 seconds, 0 frames) would reflect the same real world time. These five cameras were placed at various angles to provide dimensional data for Shawn Becker's, a Ph.D. candidate at the MIT Media Lab, image processing routines.

In the week following the production, betacam footage was transferred to the D1 tape format. Selected portions of the D1 tapes were then digitized into Y and BR NTSC encoded DAT files. These files were then converted into RGB DAT files using a utility written by Stefan Agamanolis.

3.12 Post-Production Tools

For SAR we built XmDAT, a visual annotation program that allows the user to specify key elements in a RGB DAT file as well as important historical information such as time codes, sequence names, creator, et cetera. The element annotation was used in conjunction with a chromakey program to produce RGBA images. These images are then cropped and are ready to use in SAR.

3.13 Shot Composer

The Shot Composer is a three dimensional environment that uses the images created through production and post-production. The video objects are then postitioned and moved in the environment according to environment reconstruction methods developed by Shawn Becker. Because the processes thus far have produced a virtual space of set and characters that is free form and devoid of cinematic notions, the director must create “shots” post-facto. In order to create a shot the director selects key frames by using the Composer, a tool that has some primitive navigation mechanisms which allow the user to view the scene from different positions and angles. The user can create "shots" by recording key frames with the record utilities included. These shots have associated with them a start time, when the first frame is recorded, and an ending time. These times like all time in the system is specified in SMPTE timecode format with a maximum time of 23:59:59:29. Movement between key frames is handled using bi-cubic interpolation methods.

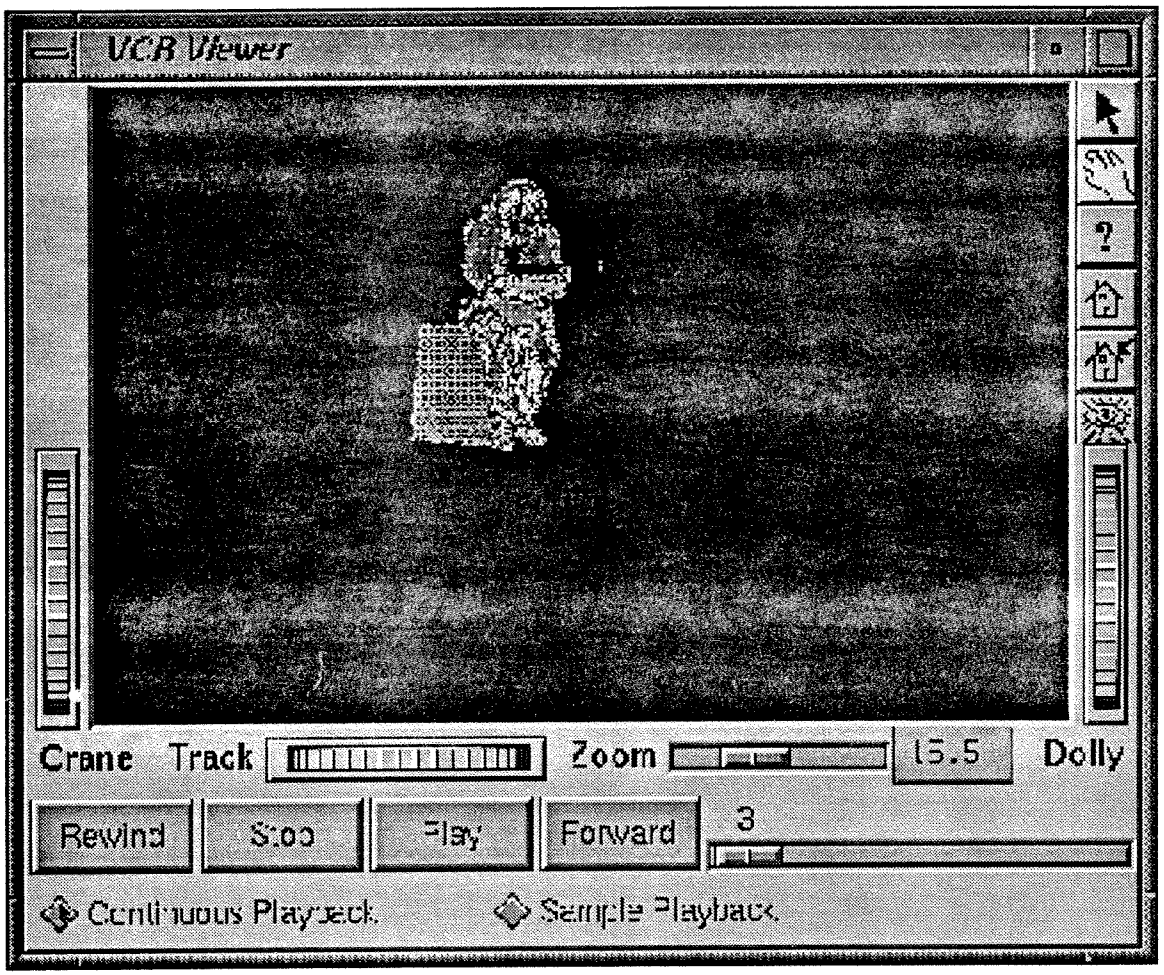

Figure 4. - Shot Composer and Player

4. Sequence Orchestrator

Thus far in the SAR tool kit we have tools for production, and post-production and a tool for creating structured video sequences. However, we still need a tool for specifying and constructing a dynamic playback of a total scene. The Sequence Orchestrator tries to fill this role, but it is important to realize that the Orchestrator is not an attempt at an editor that edits by understanding narrative flow. SAR's Sequence Orchestrator instead attempts to use some basic Hollywood conventions along with human input to construct it's scenes.

As can be seen below in Figure 5 the Shot Orchestrator is set up as a mutli-track timeline. Cameras are placed into these tracks according to the setup specified by the user in the Shot Composer. All the basic information is

displayed at each camera, the name, and the start and end times. The camera can be quickly compared to it's neighbors to determine what cameras are present. Under each camera are edit segments. These segments are specified by the user, an editor or a director. Each of these segments marks times when the camera is useful to the system, whether for narrative or coverage.

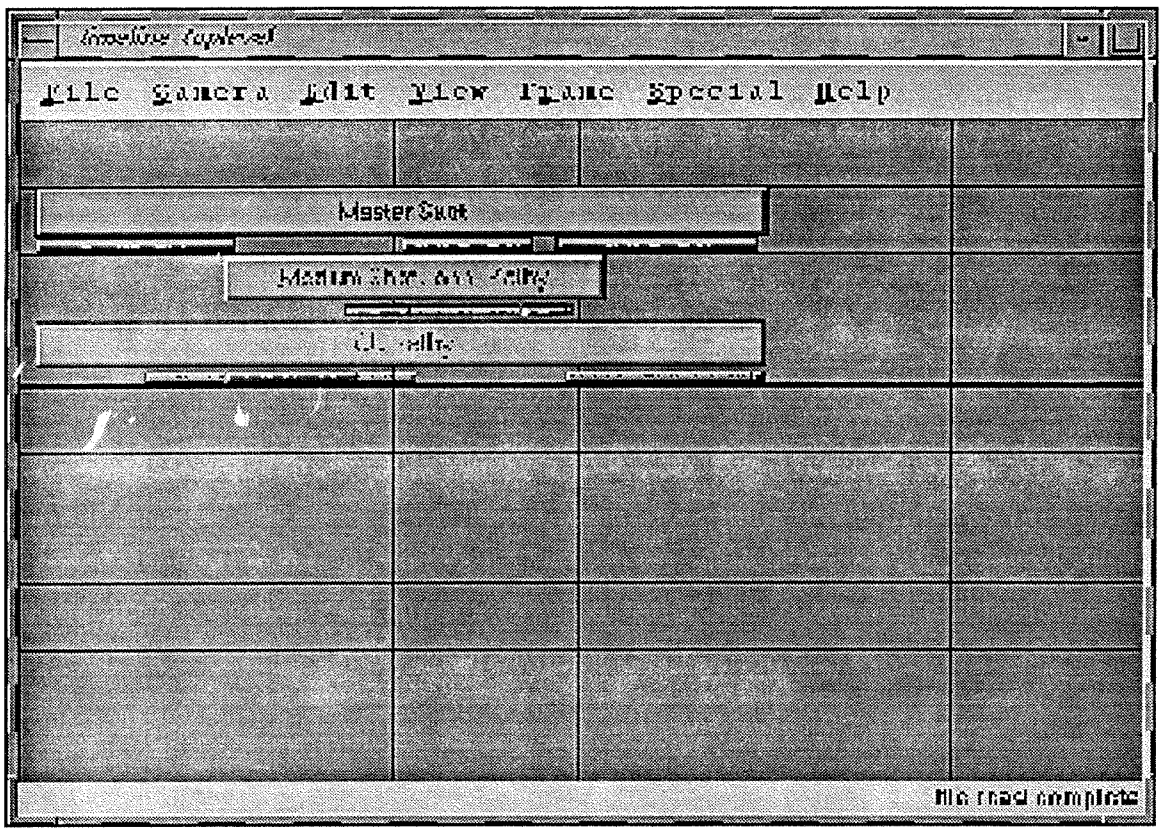

Figure 5. Sequence Orchestrator.
Cameras are the large gray bars
Edit Validity Segments are small gray bars
Edits chosen by the computer are red

Since we wanted to approach the idea of producing different scenes for different perspectives, these edit segments are perspective dependent. SAR recognizes the following perspectives, Generic, Neutral, Kathy, and John. A more complete discussion of cinematic perspective is contained in David Tamés' Master's Thesis, but for the scope of this paper it suffices to view perspective as the base narrative colored by a character's emotions and biases.

Once the Orchestrator contains data on camera duration and camera validity, editing is possible. As the reader will come to see these cameras

should be entered into the system in an order corresponding to the least specific shot at the top to the most specific at the bottom (See Figure 5 for an example). The primitive editing algorithm implemented is dependent on this ordering.

The alogrithm returns a series of EDL's (Editing Decision Lists) covering all the possible edit situations. These are lists that contain information on when to cut from one shot to another. Our algorithm for editing simply selects shots in order from time to time,choosing what kind of cut to make. Currently this decision is made primarily on the size of the viewer's screen. It is by this method that we propose to circumvent the cropping done to films today.

The program checks for the size of the display and then chooses an appropriate "shot". According to our rubric on a larger screen you want to choose larger shots since they reveal more of the cinematic space and have enough resolution to allow noticeable detail. Conversely on smaller screens those wide shots would render detailed elements such as facial expression so small as to be indistinguishable, therefore we choose tighter shots to compensate. Edits are then made wherever a new valid shot appears that fits our rubric better. The reader will notice that this algorithm is in a fairly primitive state, but several factors allow SAR to actually be a powerful automated editor.

The Algorithm operates as follows:

Algorithm: (variables are in italics)

```
time = 00:00:00:00
end_time = 00:03:00:00; (this time is standard in our system)
while time != end_time do
        find all cameras with valid edit segments at time
        if NONE signal error
        else if FRAME_LARGE then
                choose least specific shot at time
        else
                choose most specific shot at time.
        time = time + 00:00:00:01
end of while
```

The algorithm cycles through each frame of a sequence and looks at each of the possible choices at that frame. Remember that a valid choice is a camera that has an active edit segment at that frame in the scene.

The rubric places no burden of knowledge upon the computer. In order for it to construct a scene it simply needs a list of cameras and validities. The computer does not need narrative intelligence or vision recognition built into it. Instead of depending upon artificial knowledge, we tap the natural source of knowledge sitting at the computer, the director or editor. By arranging cameras in certain orders and combinations of edit validity segments directors can communicate the original intent inherent in their films.

As was stated earlier, MSL (Master Scene Language) is not only a language for specifying framingin production, but suggests editing strategies as well. The most common editing template is used for dialogue scenes. These scenes start on an establishing shot, a wider shot to reveal the space, and then at a breakpoint, typically the time when dialogue begins, editing switches between two cameras in what is know as "crosscutting." There is an ending breakpoint where the scene switches back to an establishing shot to facilitate editing to the next scene. In SAR the user can enter the two breakpoint times and two cameras and a maximum duration of a single shot. SAR will follow its normal rubric as described above. When a breakpoint is reached, however, SAR will switch to the first of a pair of specified cameras that is valid and use it, editing to the other camera if the maximum duration is exceeded or validity on the first camera runs out. If neither camera is valid at a given time, SAR simply uses its normal rubric.

Once the computer has constructed a scene we feed the editing list to the Shot Composer which will proceed to play the entire scene from the beginning of the editing list.

5. Extensions

SAR is by no means a complete system and there is much room for continued research. Below are several suggestions for areas of further development.

Navigation : Navigation is the weakest link of the current Shot Composer model. Trying to place a camera is nothing short of a wrestling match with a mouse and the dials provided by Inventor. The problem of camera navigation

was tackled in the Moviemaker's Workspace project and the concepts developed there could be applied in a useful fashion in SAR.

Rubric : It is fairly obvious that the algorithm for the shot selection could use some efficiency tuning in addition to some smarter rules for selections. As an example there may be a period of time when you have a situation of an edit like this : CAM 2 (30 seconds long), CAM 1 (.5 seconds long), CAM 2 (30 seconds long). In a situation like this it would be desirable to stick with CAM 2 instead of switching to CAM 1 as is dictated by the current rubric. Additional fields such as a narrative ranking on edit segments also would ease selection of detail shots in a sequence, as in a shot of a tea cup that a director wants popping up time and time again no matter what system you're on.

MSL Templating : An easier interface for the MSL templating for dialogue would be welcome. In addition other standard MSL templates or even interpersonal language IP interactions would be interesting to support.

Sound : Currently SAR exists, as did Moviemaker's Workspace, in the world of silent cinema. Sound poses additional challenges in maintaining smooth flow across edits.

Isis Scripting : Isis is a scripting language developed for use with Cheops by Stefan Agamanolis. Isis controls the playback of a structured video environment. If SAR produced Isis scripts then instead of playing back in the inefficient mechanism of the Shot Composer we could demonstrate our concepts on Cheops itself.

6. Conclusions

Some Assembly Required (SAR) offers an interesting approach to the lack of directional control in the pan and scan methods commonly employed today in the film business. However, in its current implementation it suggests an approach, but does not provide a complete solution. As it is currently implemented it is an at best impractical system. The entire process of digitizing and converting and annotating and finally cropping is time consuming and tedious. If in the future someone should extend the concepts of the Shot Orchestrator such a tool could inlcude knowledge of narrative and cinema that extends beyond the simple templates provided by MSL then perhaps SAR and structured video could become a viable communications method. But structured

video is a concept in its infancy and is not fully understood in terms of its impact on film.

Will a system like SAR ever be used? The answer lies in whether SAR solves or exacerbates the problems in film and media. It is easy to picture a director either happy to see the visual information and intent of his film being preserved, but it is just as easy to see a director horrified at the automatic mutilation of his master piece film.

7. Bibliography

Agamanolis, Stefan. "High-level scripting environments for interactive multimedia systems." M.S. Thesis, Massachusetts Institute of Technology, 1996.

Bove, V. Michael. Jr. and Watlington, John. "Cheops: A Reconfigurable Data-Flow System for Video Processing," *IEEE Transactions on Circuits and Systems for Video Technology*, 5(2), 1995.

Drucker, Steven Mark. "Intelligent Camera Control for Graphical Environments." Ph.D. Thesis, Massachusetts Institute of Technology, 1995.

Evans, Ryan. "Log Boy meets Filter Girl." M.S. Thesis, Massachusetts Institute of Technology, 1993.

Galyean, Tinsley. "Narrative Guidance of Interactivity." Ph.D. Thesis, Massachusetts Institute of Technology, 1995.

Gilman, Charlotte Perkins. *The Yellow Wallpaper*. Thomas L. Erskine and Connie L. Richards, eds. New Brunswick, New Jersey: Rutgers University Press, 1993.

Higgins, Scott. "The Moviemaker's Workspace: Towards a 3D Environment for Previsualization." M.S. Thesis, Massachusetts Institute of Technology, 1994.

Morgenroth, Lee. "Movies, Talkies, Thinkies: An Experimental form of Interactive Cinema." M.S. Thesis, Massachusetts Institute of Technology, 1992.

Richards, Ron. *A Director's Method for Film and Television*. Boston, Massachusetts: Focal Press, 1992.

Tamés, David. "Some Assembly Required: Cinematic Knowledge-Based Reconstruction of Structured Video Sequences." M.S. Thesis, Massachusetts Institute of Technology, 1996.

Winston, Patrick Henry. *Artificial Intelligence*. Reading, Massachusetts: Addison-Wesley, 1992.