Towards Virtual Videography

Michael Gleicher
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706
gleicher@cs.wisc.edu

James Masanz
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706
aranduil@cs.wisc.edu

ABSTRACT

Videographers have developed an art of conveying events in video. Through choices made in cinematography, editing, and post-processing, effective video presentations can be created from events recorded with little or no intrusion. In this paper, we explore systems that bring videography to situations where cost or time issues preclude application of the art. Our goal is to develop *virtual videography,* that is, systems that can help automate the process of creating an effective video presentation from given footage. In this paper, we discuss how virtual videography systems can be constructed by combining image-based rendering to synthetically generate shots with image understanding to help choose what should be shown to the viewer. To this, visual effects can be added to enhance the presentation, lessening the degradation caused by the medium.

Categories and Subject Descriptors

I.3.3 [**Computer Graphics**]: Picture/Image Generation; I.4.6 [**Image Processing**]: Segmentation; I.4.9 [**Image Processing**]: Applications

Keywords

Educational technology, videography, automatic presentation

1. INTRODUCTION

Portraying an event, such as a performance, lecture, or sporting event, in video is an art form practiced at many levels. While the experience of watching an event on screen is different from being there, well produced video can provide a differently effective experience for the viewer. Skillfully produced video can compensate for the limitations of the medium by exploiting the power of cinematic media to manipulate time and space and artificially enhance images, as well as to avoid some of the problems with attending a real event, such as being restricted to a given seat.

Videographers have a number of tools at their disposal for capturing and portraying an event. Multiple cameras, with various lenses, provide multiple viewpoints. The raw "footage" provided by the cameras is edited together to point the viewer's attention to where it is most needed and to control the timing of the presentation: compressing time by skipping over unimportant segments, or dilating it by replay or slow-motion.

The hardware demands of videography are a barrier to its use. Equally barring is the skill required at each phase of the process. Skilled camera operators are needed to capture the significant events with movements that will not induce motion sickness. Skilled editors and directors are needed to choose which shots should be used and when. These two phases are tightly coupled: the director often guides the cameras to insure needed footage is available, but ultimately must choose among what is provided, sometimes augmented with archived footage or synthetic images.

The challenge of videography is related to, but different than, traditional cinematography. The videographer has limited control over the events that are being filmed: there is no mise-en-scène [5]. Ideally, the videographer would unobtrusively observe and record what happens, although this ideal is sometimes compromised (for example by an intrusive wedding photographer).

Unfortunately, many applications do not afford the use of videography. Often, cost and intrusiveness considerations limit the number of cameras and their mobility. Cost and availability concerns often preclude the use of skilled practitioners, both during filming and production.

2. VIRTUAL VIDEOGRAPHY

Our goal is to construct a system and methodology for recording events with minimal intrusion, and to produce effective video from this footage in as automated a manner as possible.

Consider the task of creating video presentations from class lectures. In such a setting, cost and intrusion considerations preclude the use of more than a small number of non-mobile cameras. We would not expect to either recreate the experience of being in the class, or the video that might have been created were the whole event designed as a video. We do not want to affect either the instructor's presentation, nor the experience of the students in the class.

We are beginning by focusing on off-line systems. Some applications, such as live broadcast, require real-time, online systems. No application would be hurt by a system having an on-line capability, and such an on-line system has been explored by Bianchi [4]; however, in an off-line system we have certain advantages:

- Looking ahead in time can help us anticipate the action. In an on-line application, more knowledge of the event is required to help predict what will happen if unpleasant surprises are to be avoided.

- By looking at durations of the presentation simultaneously, we can better enforce temporal constraints, for example avoiding jittering and adhering to the 180 degree rule [5].

- Information from previous or future frames can be used to create special effects.

- The system need not operate in real-time.

Our target medium is to create "standard" video; a linear presentation. While interactivity offers potential for a novel presentation medium, we prefer to limit ourselves to a more traditional medium where presentation techniques are better understood. Much of the art in cinematography is in guiding the viewer's attention. Determining how to employ the existing art is challenging enough.

2.1 An Example Problem

We have chosen a specific, limited domain in which to explore virtual videography: medium sized classes given by a single lecturer in "chalkboard" style. This domain shows the need for virtual videography: while there is clearly value in making such material available to those unable to attend the lecture, cost considerations preclude the use of a professional video staff. Placing one or two static cameras in the back of the lecture hall is practical. Not surprisingly, these static camera videos are considerably less interesting to watch than the original lecture itself.

Our goal is to be minimally intrusive, not requiring the lecturer to change the presentation at all. Our view is that we are recording an event, not creating a different one. The presentation is really meant for the students in the class, and the instructor should be free to teach using whatever method they have honed for best communicating in this setting. We will use this simple domain as a running example through the paper.

LectureBrowser [15] also aims to non-intrusively record university lectures and create video presentations. They synchronize the observed lectures with display of other digital media, and rely on a known lecture format, tracking hardware, and apply cut-only editing between two fixed views.

The Classroom 2000 Project [1] makes a record not only of the lecturer, but also of the lecturer's notes and notes taken by students. The project does not aim to be completely non-intrusive; it aims to capture the entire event, including a record of the students notes.

3. SUBPROBLEMS

Producing "real" videography requires a team, or at least a multi-talented practitioner. Similarly, a virtual videography system requires a range of components. In this section, we survey these components and the issues that must be addressed.

Each component of a virtual videography system is an open-ended research topic in its own right. However, all afford a number of simpler solutions that can be constructed today without major extensions to the state of the art. In each of the following subsections, we note not only the potential for future systems and research directions, but also our initial experiments in the example domain.

3.1 Data Sufficiency

Given the fixed limited set of source images, we must first ask whether or not there is enough information to create the desired result. This problem is inherent in off-line production, not a consequence of virtual videography. A human editor faces this same problem when presented with the same raw footage.

The sufficiency questions arise at all levels. For instance, if there is insufficient information to see some detail, then there is simply no way to show that to the viewer. These sufficiency questions can be difficult to determine: what may be unreadable at first might be curable using image enhancement, or by combining elements from several sources. At a higher level, if a topic is not discussed in a presentation, it is unlikely that it can be explained in the resulting video.

Sufficiency issues lead to two general questions: How does the recorder of the event determine if there will be enough information in the "sampling" being recorded to create a good result? And how do we best use these bits to communicate a desired message? For example, when we record an event, can we know if two cameras are sufficient? If so where should we place them? And, given the output from these cameras for a given performance, how do we best use their images to convey the presentation? In our example domain, we have little control over the amount of data that we can obtain, and focus on the last question.

3.2 Image Understanding

More understanding about what is occurring in the source video footage enables more informed choices in how to utilize it. A virtual videographer must use computer vision to interpret its footage, or rely on manual intervention. A virtual videographer's needs are standard vision issues, such as person tracking and gesture recognition.

The vision task is simplified for the virtual videography application because precise results are not necessary. Initially we may further simplify the problem by limiting the scope of our application to a constrained and structured environment, allowing user intervention, and demanding only coarse grained information from the vision system. In our example domain, the static camera and knowledge that the only moving object is the presenter allows simple techniques, such as change detection and skin-color classifiers[9] to be effective.

To relax these restrictions, more sophisticated image comprehension allows for the automatic identification and interpretation of the action. For example, we might not only identify that there is a person gesturing, but that they are pointing at a particular object. Understanding where the action is taking place or where the attention of the people in the scene is focused suggests where the videographer should direct the viewer's attention.

3.3 Computational Cinematography

Much of the power of cinematography comes from the

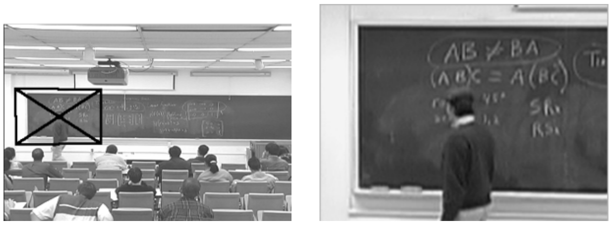

Figure 1: Simulated medium shot.

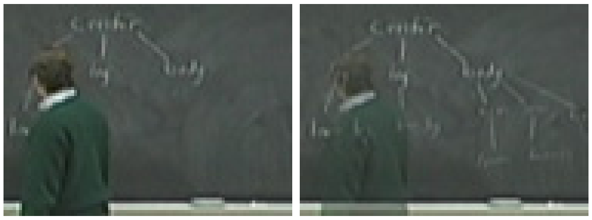

Figure 2: A Compositing Effect.

ability to control the viewpoint. Through this control, the limited portal of the screen can be expanded through motion, as well as focusing the viewer's attention [12]. A virtual videography system chooses what viewpoints to show. Care must be taken to not only properly guide the viewer's attention, but also to not confuse them (unless it is intentional). When done correctly, such continuity-editing can seamlessly guide the viewer through time and space. Cinematography and editing is an art form unto itself.

In computer graphics, there have been various attempts to codify the art of cinematography. Karp and Feiner [11] use planning techniques to make cinematographic decisions based on knowledge of communicative goals, while He et. al. [8] explore automating cinematography in the context of animated conversations.

For creating video presentations, the choices are more limited. Bianchi [4] shows how a simple set of heuristics can be effective for videography, while LectureBrowser [15] gives an even smaller and simpler set to make cuts between a wide and close-up. In our example domain, we plan to mimic these heuristics, and extend them using our ability to look forward and backward in time to insure sufficient continuity and variety. Creating a presentation using heuristics is in contrast to the use of authoring tools such as MAD [3] that leave editing decisions to the author, and systems that give the viewer choices during playback, such as STREAMS [6].

3.4 Image-Based Rendering

We define the virtual videography problem as beginning with a set of source images. However, this set might be smaller than would be desirable, especially since no director was available to guide the shooting. This would mean that there would be a very limited selection of viewpoints. However, graphics and vision researchers have been exploring methods for generating novel viewpoints based on an initial set. While general solutions to the view interpolation are on the horizon, interesting methods have emerged for various special cases, for example those of Manning and Dyer [14], Seitz [16], and Avidan and Sashua [2].

A very simple form of novel view generation can be created by assuming that the camera only zooms and rotates around its optical center. The mapping between any two images taken from a camera with a fixed center is a projective deformation [17]. Therefore, panning and zooming can be implemented using a projective warp, as shown in Figure 1, where a dynamic close view has been created by re-sampling a longer, static one.

3.5 Shot Generation (Camera Operation)

Image-based rendering addresses the problem of creating a view, however, we still must control it. In analog to real videography, image-based rendering creates the camera, however we still need to implement the cameraman. Camera operation must consider both the spatial aspects, how the "individual pictures" look, as well as the temporal aspects, to create motions that do not confuse or sicken the viewer.

Ultimately, a virtual videography system may encode heuristics that define the art of photography and cinematography. This has been explored in a constraint-based framework by Drucker [7] in the context of 3D virtual environments. Initially, our virtual videography experiments use simple algorithms to frame important elements, and we use filtering to avoid the generation of jittery motions. Implementing the filtering by fitting known good movement patterns, such as ease-in/out, will further improve the results.

3.6 Special Effects

Applying special effects is another form of shot creation. There are a wide range of effects that can be put to use: super-imposition, transparency, transitions, titling, picture-in-picture, etc. In our example domain, we imagine highlighting what a presenter points to. Traditional methods of emphasis, such as pointing, often obscure the very thing to be emphasized. With special effects we have the opportunity to emphasize something without obscuring it further.

Figure 2 is an example of another effect useful in our examples. In the original frame (left), the instructor obscures the partially drawn diagram. By combining this image with a later one, the obscured text is revealed, as is a sense of where the instructor is going. The right frame was constructed by overlaying a partially dissolved copy of the original frame over a frame taken from later in the video when the writing on the board is complete.

The inclusion of special effects make other aspects of the videography problem more difficult. The system must determine when and where to use them and what source footage is necessary to best generate them. There is also the question of whether such visual devices are effective, or are they confusing and distracting. Avoiding these latter problems will require developing ways to cue the viewer to what is happening.

4. INITIAL EXPERIMENTS

As stated earlier, a virtual videography system has a number of components, each with an open research agenda. Our approach to virtual videography is to aim for building a complete end-to-end system, with engineering "place-holders" for each component. Once such a system is demonstrated, we can further address each component in the context of a complete application. In this section, we describe our initial experiments and prototype.

For our initial explorations, we recorded an entire semester of lectures in an undergraduate course using DV camcorders.

Generally, a single static camera placed in the rear of the room was used, although a limited number of lectures were filmed with two cameras.

4.1 Proof of Concept

Our first efforts aim to show that there is in fact sufficient information in our source materials to create our targets. Given the extremely limited source material, is it possible to produce video of the sort we aim for? If not, how much more should we sample the lecture, or what concessions to invasiveness should be made? To experiment with this, we have chosen a "Wizard of Oz" prototyping approach [13] [10] where a user manually does the process envisioned by the final system. We have done this by attempting to produce video using commercial video production tools.

Some findings:

- Standard production software is not especially suited to our task.
- Manipulation of audio is not required, despite the moving viewpoint.
- Care must be taken with placement of the camera to make sure the chalkboard is readable on tape.

4.2 Initial Prototype

Our initial virtual videography system is designed so we can construct a working system as quickly as possible to explore the ideas, yet we can easily expand and improve it.

Ideally, the system will do a good enough job that human collaboration will be unnecessary when the expectations for the presentation are not too high. Initially, we rely on user interaction to compensate for simplified pieces of the system.

Our prototype is implemented on Windows NT workstations using our PyVideo Toolkit which relies on Video for Windows, Python, and the FlTk interface toolkit. Our initial experience shows that some very simple methods can produce interesting results, however many aspects of the problem require further exploration.

Acknowledgments

Rob Iverson implemented the majority of the PyVideo toolkit. This research was supported by NSF Career award CCR-9984506, a grant from Microsoft Research, and a hardware donation from Intel.

5. REFERENCES

[1] G. Abowd, C. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney, and M. Tani. Teaching and learning as multimedia authoring: The classroom 2000 project. In *ACM Multimedia '96*, 1996.

[2] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *CVPR97*, pages 1034–1040, 1997.

[3] R. Beacker, A. Rosenthal, N. Friedlander, E. Smith, and a Cohen. A multimedia system for authoring motion pictures. In *ACM Multimedia '96*, 1996.

[4] M. Bianchi. Autoauditorium: a fully automatic, multi-camera system to televise auditorium presentations, 1998.

[5] David Bordwell and Kristin Thompson. *Film Art: An Introduction*. The McGraw-Hill Companies, Inc., 1997.

[6] G. Cruz and R. Hill. Capturing and playing multimedia events with streams. In *ACM Multimedia '94*, 1994.

[7] S. Drucker and D. Zeltzer. Camdroid: A system for implementing intelligent camera control. *1995 Symposium on Interactive 3D Graphics*, pages 139–144, April 1995.

[8] L. He, M. Cohen, and D. Salesin. The virtual cinematographer: A paradigm for automatic real-time camera control and directing. *Proceedings of SIGGRAPH 96*, pages 217–224, August 1996.

[9] M. Hunke and A. Waibel. Face locating and tracking for human-computer interaction, 1994.

[10] Todd Hovanyecz John D. Gould, John Conti. Composing letters with a simulated listening typewriter non-traditional interactive modes. *Proceedings of Human Factors in Computer Systems*, pages 367–370, 1982.

[11] P. Karp and S. Feiner. Automated presentation planning of animation using task decomposition with heuristic reasoning. *Graphics Interface '93*, pages 118–127, May 1993.

[12] S. Katz. *Film Directing Shot by Shot: Visualizing from Concept to Screen*. Michael Wiese Productions, 1991.

[13] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Office Information Systems*, 2(1):26–41, 1984.

[14] R. Manning and C. Dyer. *Confluence of Computer Vision and Computer Graphics*, chapter Dynamic View Interpolation without Affine Reconstruction. Kluwer, 2000.

[15] Sugata Mukhopadhyay and Brian Smith. Passive capture and structuring of lectures. In *ACM Conference on Multimedia*, 1999.

[16] S. M. Seitz. *Image-Based Transformation of Viewpoint and Scene Appearance*. PhD thesis, University of Wisconsin - Madison, October 1997.

[17] R. Szeliski. Image mosaicing for tele-reality applications. In *WACV94*, pages 44–53, 1994.