To Appear in Proceedings SmartGraphics.

A Framework for Virtual Videography

Michael L. Gleicher Rachel M. Heck Michael N. Wallick

Department of Computer Sciences
University of Wisconsin - Madison
Madison, WI 53706
{gleicher, heckr, michaelw}@cs.wisc.edu

Abstract

There are a significant number of events that happen on a regular basis that would be worth preserving on video but for which it is impractical to use traditional video production methods. In this paper, we describe one possible way to inexpensively and unobtrusively capture and produce video in a classroom lecture environment. We discuss the importance of cinematic principles in the lecture video domain and describe guidelines that should be followed when capturing a lecture. We continue by surveying the tools provided by computer vision and computer graphics that allow us to determine syntactic information about images. Finally, we describe a way to combine these tools to create a framework for a *Virtual Videography* system, one that can automatically generate production quality video. This framework is based on the creation of *region objects*, a semantically related region of video, despite the fact that we can reliably only gather syntactic information.

1 Introduction

Many common events are worthy of recording for later viewing. As the price of video equipment and storage decreases, capturing these events becomes increasingly practical. Unfortunately, merely capturing an event on video does not make the video watchable. Creating effective video is an art requiring a wide range of talents. This complexity makes video inaccessible for applications that cannot afford a skilled production crew or the intrusion of such a crew on the event being recorded. Our goal is to lower these barriers, making it easier to produce videos from events. We propose that the observations of unattended, non-intrusive, inexpensive sensors, such as statically placed cameras, can be post-processed to create video with sufficient production value. We call such an approach *Virtual Videography* [8] as it must simulate many of the aspects of video production.

In the century since the first movies, filmmakers have developed an art of portraying events through moving images [4, 3]. As videographers have adapted this art to the "small screen," they have refined it and developed approaches for portraying even simple events, such as newscasts and speeches [15]. Many aspects of this art, including the use of cinematography, composition, editing, and even visual effects, are important for more than entertainment value. Good video enhances the message and reduces the burden of comprehension on the viewer.

Automating the production of video that can effectively communicate requires application of various aspects of the filmmaker's craft. Through proper editing and post-processing, the viewer can be lead through an event, greatly increasing the likelihood of comprehension. A system that can automatically create such video requires some knowledge of filmmaking principles. In addition, it needs to have some understanding of what is happening in the event (so it can make effective choices) and have the ability to synthesize the necessary images. While constructing such a system is a daunting task, we believe it is possible to do in some restricted but valuable domains.

In this paper, we discuss the goals and issues of Virtual Videography and the ways in which it might be realized using today's technology. We begin by describing our example domain, the classroom lecture, and discussing how the art of videography is useful in this domain. We then examine the requirements for a viable Virtual Videography system and survey the technologies that are available for its construction. Using these basic elements, we introduce the notion of *region objects*, or a region in video that is semantically related, as determined by spatial or temporal proximity. These region objects provide a framework for both design and implementation of a Virtual Videography system.

We emphasize that we have not yet constructed a Virtual Videography system. The examples shown are either "Wizard of Oz" prototypes [14, 11], where we simulate the results of the proposed systems using manual operation of current tools, or disconnected prototypes of individual components. Our goal in this paper is to introduce issues in constructing a Virtual Videography system and to give evidence as to why we believe it is plausible to create one.

2 Video Production and the Classroom

Cinematic media, such as film and video, are limiting in that they provide the viewer with a small, rectangular window on the world being portrayed. While watching a film or video, the viewer cedes control of their viewpoint to the filmmaker. Digital video technologies offer the potential for new interactive media that return some of this control back to the viewer [17]. However, the success of the traditional media has come from this transference of control: by ceding control, the viewer can be guided through a new story by the filmmaker. Over the century since the invention of film, filmmak-

ers have developed an art where they use this control effectively to deliver their message [3]. Discussion of the art of filmmaking most often comes in the context of entertainment [4]. When cinematography or editing is mentioned, we quickly think of feature films where aesthetic quality is clearly important for our enjoyment. However, the art of filmmaking is applicable, and arguably crucial, in any attempt to portray the world through the limited portal of the screen. By making effective use of filmmaking principles, video becomes a far better communicative tool. By ignoring these principles, we give up the opportunity to make our videos better.

One view of the art of filmmaking is that it compresses or expands time and space to fit within the screen and duration of the film. This is most obvious in epic feature films where in the course of two hours, we are shown a story that may span many years and many miles. Through careful use of cinematography and editing, the filmmaker can fit a large story into the small space-time box of the movie in a way that seems natural to the viewer. Cinematic principles help prevent the viewer from being disoriented (unless the filmmaker wants them to be).

2.1 The Classroom Video Domain

The portrayal of a classroom lecture in video may not seem to have much in common with the movie epic, but it also requires the application of the filmmaker's craft to compress time and space onto the video screen [15]. Many of the same principles that apply to film also apply here, although possibly in more subtle ways. We have chosen classroom lecture video as the primary domain for our explorations into Virtual Videography.

By a "classroom lecture video," we literally mean a video record of the event of a classroom lecture. That is, we want to record a real lecture, given for the purpose of the "live" audience. Even more specifically, we focus on chalkboard lectures, as we have abundant and immediate access to these types of lectures. In addition, chalkboard lectures offer an element of complexity that is not found in PowerPoint or slide lectures; we do not automatically know when the focus of the lecture changes as we do with a PowerPoint or slide lecture. Within this domain, we work under the assumption that the event is for the audience. For example, a university class lecture is for the students in the class. If a video can be produced from a lecture, it is an added benefit.

We should emphasize that the event observing model of video production is quite different from the typical ways in which videos are produced. At the beginning of the 20th century, early filmmakers learned that simply recording a theater production did not lead to effective films [3]. Usually, "events" are purposely staged in a way that presents good video. If our primary goal was to create the most effective instructional video, we could carefully plan a production to achieve this goal. Instead, our goal is to gain additional value from events that are already occurring. The videos obtained can be useful for later review or for students unable to attend, but this is not the focus of the lecture itself.

By shifting our focus to event record video, we acknowledge that the quality we will achieve cannot compete with video created for its own sake. While the quality standard might be lower, our challenges come from the fact that we must avoid altering the event and that we must keep the cost low (if we were willing to pay a higher cost, we could have done a separate video production).

This focus leads to several restrictions on what we can do:

Unobtrusive, Passive Capture of Events: It is important that the capture process be unobtrusive and passive. The participants of the event should not have to be aware of the capturing process, beyond giving their consent. This means that the cameras and other recording devices should be out of the way. Some minor adjustments, such as asking the lecturer to wear a lavalier microphone, may not be unreasonable.

Lecture Style Does Not Change: It is equally important that the lecturer not have to change his or her style to fit into the video system. Many lecturers have spent several years perfecting their teaching style. It would be a detriment to the lecturer and students to force a change in style.

Video and Audio Sensors Should Be Inexpensive: We would like the equipment to be low cost. Reducing costs increases the ease of deployment.

2.2 Cinematic Principles for Lecture Video

Even in the restrictive domain of lecture video, cinematic principles are critical to making video that communicates effectively. While the event being portrayed may not be expansive, it still must be compressed in space (and potentially time) to fit within the bounds of the video screen.

The medium of a live lecture is clearly different than that of video. Therefore, the problem of portraying a lecture on video is one of translation. Working backwards, we can see that the properties of good video all have analogs in the lecture domain. By paying attention to the properties of good video, we can preserve the properties of good lectures:

Guide the Viewers Attention: In both media, the presenter (videographer or lecturer) must properly guide the viewers attention. In the classroom setting, the lecturer can direct a student's attention through gesture or description. The videographer has more responsibility over the viewer's attention, since they control what the viewer sees. But the videographer also has additional mechanisms for guiding the viewer, such as close-up shots, zooms, and special effects (e.g. lighting up an object of interest).

Make Effective Use of Time and Space: The main space in which a lecturer presents is provided by the chalkboard. A lecturer must make effective use of this space to present material in a comprehensible order. While such layout may not always be carefully planned, use of the entire chalkboard space is common. This can provide students with maximum context for the lecture and allow professors to return to previous parts of the lecture as needed.

The space of the video screen is quite different from the chalkboard in size, space, and resolution. It is both smaller, in that far less information can be shown at any given time, yet infinite, in that it can easily be changed to display any portion of anything shown during the lecture.

The timing and pacing of a lecture is clearly part of the presenter's art. However, some of the timing may come from factors that are not required in video. For example, pauses for erasing portions of the board may provide little pedagogical value. When applied carefully, changing the timing during translation to video by cutting silent or identifiably uninteresting portions or by altering the speed of other portions may enhance the value of a video lecture by tightening its message.

Maintain Visual Interest: Because of the limitation of the small, fixed screen, cinematic media must use a variety of shots to

provide visual interest. While this is obviously important for keeping the viewer interested and engaged, visual interest also provides cues to the viewer about what is important and what is not.

Communicate the Message: Generally, a lecturer is trying to convey some material to the audience. It is important that this same message is also conveyed to the audience of the video.

2.3 Lecture Capture

The guidelines we have put in place for lecture capture dictate the use of passive capture. Given the decreasing cost of video cameras, it is practical to mount them unobtrusively within a lecture room. The most obvious placement of a camera is at the rear of the room, providing a view of the entire lecture area. For our experiments, we deployed multiple video cameras placed close together to provide a single, higher-resolution image.

One potential extension is to mount the video cameras on pan-tilt mounts controlled by an automatic tracking system. Such an approach is used to provide a close-up image of the presenter in systems developed by Foveal Systems [2, 6] and Microsoft [18]. Such devices are expensive and require access to a computer and specialized software during capture. Instead, we choose to use simpler sensors and to create other images by post-processing.

For a lecture, the audio is a critical component. Capturing the audio effectively is difficult without intruding on the participants. Ceiling mounted arrays of microphones can provide coverage over a large area. However, these microphones are expensive and difficult to install. For capturing the lecturer, a lavalier microphone requires some intrusion but provides significantly better audio quality than other devices. Combining the lavalier microphone with a ceiling mounted array provides good quality audio of the lecturer as well as some audio of the audience.

Digital and video imagery provide finite resolution, which becomes an issue if we need to expand an image to simulate a zooming operation. While we have tried to use the best available video recording formats (mini-DV), the nature of video is limiting. We have tried to address this by treating multiple cameras as a single, high resolution image (through mosaicing). We have also explored the use of a digital still camera to provide high-resolution data. While the NTSC video format provides (at most) $720x480$ resolution, digital still cameras can provide many times that. However, video cameras are capable of recording 30 frames (or 60 half-resolution fields) per second, while still cameras are much slower. Combining both video and still cameras provides a cost-effective mechanism for obtaining data sets that are high-resolution in both time and space. It does, however, provide a sensor-fusion problem.

Other recording devices can be added to provide additional information about the lecture. Some potentially interesting devices are chalk/marker board recording devices such as the Mimio [23] which tracks marker and eraser locations or "ZombieBoard" [19] which uses computer vision to scan and interpret writing on a marker board. By providing accurate position and timing information, such devices provide an interesting source of additional data that can be fused with the imagery. To date, such devices have been impractical in lecture settings because of their limited range.

2.4 Current Approaches

The utility of recording lectures without the expense or intrusion of a video production crew has inspired a number of efforts besides ours. Similar to our project, several groups of researchers have built systems that automatically produce lecture video by switching between multiple, simultaneous streams of video. Researchers at Microsoft have constructed a system that uses computer vision to select between multiple video cameras (including a tracking camera) in real time. The rules for selection are informed by discussions with professional videographers [18]. Foveal Systems' Auto Auditorium product shows that such systems can be made robust enough for commercial deployment [2, 6]. Mukhopadhyay and Smith [16] show a simpler system that switches between two cameras.

All three of these systems can provide "edited" video in real time. Our Virtual Videography system is distinguished by not being real-time. While this is a limitation, it allows us to perform more significant processing and to consider a wider range of display options. Rather than limiting ourselves to a fixed set of camera views, we can potentially synthesize novel views based on past, present, and future information. Also, all of the other systems work in the more restricted domain of slide lectures. The format of a slide lecture is more controlled, and there is a more fixed set of cues (changes of slides).

A related set of efforts attempts to add value to passively captured video by creating meta-data for search, browsing, visualization and/or summarization. The CUE-Video effort at IBM Research [22] is notable as it uses a combination of vision and signal processing techniques to address all of these questions.

Another related set of research uses cinematography knowledge to create animated presentations. For example, Drucker and Zeltzer [5] cast the camera motion planning problem as a constrained optimization, He et al. [9] choose camera motions for virtual dialog scenes, and Bares and Lester [1] automatically control cameras in virtual worlds.

3 Available Building Blocks

In order to produce a video, a videographer uses their observations and understanding of the action to determine how best to portray it on screen. At best, they have a deep understanding of the event. Unfortunately, we cannot expect an automated system to have this degree of information.

The challenges of computer vision make it difficult to provide a sense of sight to a Virtual Videography system so that it can make informed choices. Given the current state of the art in image understanding, it is unlikely that an instructor's writing on the chalkboard could be converted to text, let alone interpreted [24]. Similarly, the poor quality of the available audio and the technical nature of the dialog make it accurate speech recognition unlikely.

We expect that in the near term, it will be infeasible for a Virtual Videography system to gain true semantic knowledge of what is occurring in the lecture. Any system will be limited to purely syntactic information. That is, the system can know that there are marks made on the chalkboard but will have no way to know what these marks mean. Similarly, the system can know that the instructor is speaking and maybe even spot certain words, but it will not be able to produce a transcript or understand what is being discussed.

The computer vision problem is so challenging that even low-level syntactic information can be hard to obtain. Fortunately, the simple environment of classroom video, coupled with the use of static cameras, permits the use of several types of algorithms:

Color Segmentation: Color segmentation classifies each pixel into groups based on color information. Color segmentation

is often used to find skin in images, however it can be used to describe almost any color group. This can be used in Virtual Videography to find the chalkboard and professor throughout the video. [10, 12]

Motion Segmentation: Video can be separated into regions based on motion. In the lecture domain, since typically the only object moving is the lecturer, this may be a useful technique.

Tracking: Tracking allows us to follow a region of interest as it moves throughout the video. In particular, tracking can be used to follow the lecturer as he or she moves around during the class.

Gesture Recognition: Determining gestures made by a person is difficult in general but can be done in restrictive settings [7], such as a lecture environment [13]. While it will be virtually impossible to correctly interpret every gesture, we believe that slight imperfections will only result in a minor shift in focus.

Background Modelling: Background modelling is a statistical function for building a model of the background in the image. Generally, the background is static. In the classroom setting, the background (chalkboard) is constantly changing, so while it may be possible to build a coarse background, it may not be possible to generate a rich and detailed one.

Similar limitations arise in our attempts to synthesize new images to show in the video productions. Because we do not have high-level knowledge of what is happening in the scene, we cannot build a detailed model and therefore are restricted in what we can produce. Our image synthesis process is limited to re-using the pixels that were captured. We can use resampling (e.g. zooming or warping), filtering or other image processing, and compositing or other methods of combining multiple images. Image-based rendering may be able to provide us with a more complex recombination of images, however, we are fundamentally limited in that we cannot create information that was not captured. We might be able to find missing information elsewhere, such as filling in an obscured region by using an image from a different angle or time. However, if no image contains what we want to show, we cannot show it.

4 A Framework for Classroom Videography

The tools provided by available computer vision and graphics technology are quite low level. They provide us with only a small amount of information about what appears in the images and absolutely no information about the semantics of the scenes captured in the images. However, we believe that in the restricted domain of classroom videography, this low-level syntactic information is sufficient to do interesting videography.

In this section, we propose a scheme that builds on the provided building blocks of computer vision and graphics. We restrict ourselves to techniques that we can plausibly create using the current state of the art. The key idea is to break the chalkboard into smaller pieces that we call *region objects.* Region objects provide not only an implementation strategy for a Virtual Videography system, but also a tool for thinking about how filmmaking knowledge can be cast in a computational framework.

4.1 Segmentation and Layers

Layered video [25, 20] breaks a video stream into a number of independently moving, stacked, 2D pieces. For general scenes, robustly and accurately breaking a video into these pieces is challenging. However, in the chalkboard domain, our task is simplified. There are two layers, the instructor and the chalkboard. The chalkboard is static and easily identifiable. Using a combination of color and motion cues, such segmentation should be possible.

Creating a true layered representation requires more than simply determining which object each pixel belongs to. For objects that are partially obscured, we must determine what the object looks like in the obscured region. Because we cannot see the obscured region, we can only speculate as to its appearance, typically based on previous or future frames of video. Because the chalkboard does not move relative to the camera and we have a relatively good understanding of how the appearance of the chalkboard evolves, we can make a good guess about obscured regions by looking forward to a time at which the region is not obscured.

The segmentation process provides us with separated video streams of the audience, instructor, and the chalkboard. The audience layer does not contain much useful information and to use it would require consent of the entire audience, therefore, the audience layer can be filtered out and replaced with a video textured audience [21].

After the audience has been removed, we are left with the lecturer and the chalkboard layers. Each layer serves very different purposes and can be handled in different ways. This segmentation scheme also provides us with information about what pieces of the chalkboard are visible at any instant. This can be extremely useful when we try to fuse information from multiple sensors. For example, when analyzing the images of the still camera, we lack motion cues for doing segmentation. The fusion of sensors can provide additional views or information of various regions of the two layers.

4.2 Chalkboard Regions

Given the chalkboard, we can divide it into smaller regions related by content. Unlike traditional video layers, these *region objects* do not move, as they are objects written onto the black board. For simplicity, we consider the regions to be rectangular.

Over time, the writing on the chalkboard changes. Any particular region object may be growing (the instructor is writing it), shrinking (portions are being erased), or remaining constant. Even when the size of the region is remaining constant, the content may not. A *region object* is a spatio-temporal object that contains not only the image data, but also its temporal progression. A schematic is shown in figure 1.

For a given region object, we may have many potential sources of its image. For example, when a region object is constant, we might want to take its image from a different frame if it is currently obscured or even from a different sensing device that can get a clearer picture of it.

Ideally, the region objects would represent semantically relevant groupings, such as sentences, diagrams, or phrases. Lacking the ability to understand what the markings within regions mean automatically, we must approximate the goal with syntactic knowledge. By using spatial and temporal proximity, we can segment the chalkboard into region objects.

Unfortunately, we cannot rely on the groupings having semantic relevance. Higher-level semantic information would allow region objects to be grouped hierarchically. It is unclear that there is a purely

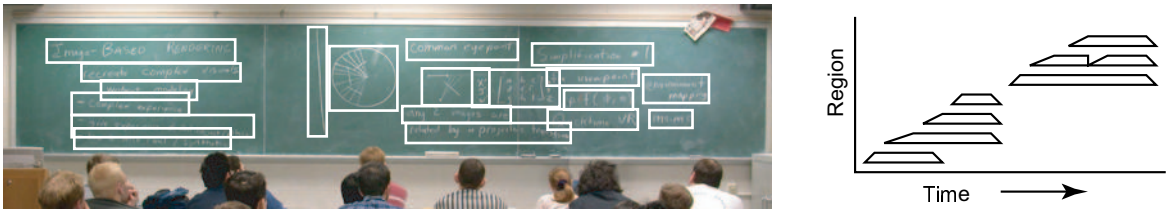

Figure 1: Two views of region data: the left shows the spatial arrangement of the regions at a given instant; the right shows their temporal evolution.

syntactic analysis that could automatically provide such groupings. Syntactic groupings of region objects, based on temporal proximity, spatial proximity, or order of viewing (see the next section) are certainly possible.

After performing region analysis, we will have developed a set of rectangles, their evolution over time, and their preferred view. This provides a platform for understanding the action of the lecture, as well as the opportunity to process the region objects independently and assemble them in response to our communicative goals.

4.3 Regions of Interest and Gestures

One of the tasks of the lecturer is to guide the audience's focus of attention. A goal of the results of Virtual Videography should be to have the video mimic the direction of focus of the lecture as closely as possible. We express this concept in the framework of region objects as having a *region of interest* object. Alternatively, we refer to one of the existing regions as being the *focus*.

At any given instant, there is a set of region objects on the chalkboard. One of them may be distinguished as the region of interest, that is, the region that is the most likely focus of the audiences attention. While there are a variety of subtle cues that the instructor may use to drive the audiences focus, we believe that a set of simple, syntactic heuristics can be applied to find this important region. For example:

- a newly created region is likely to be the focus;
- all regions should be the focus at some point during their existence;
- many of the instructors gestures, such as pointing, can be interpreted as specifying the region of interest;
- things should not stay the focus for too long, or for too little time.

Without the hierarchical understanding of the connections between region objects it is difficult to understand what the context required for understanding the object of interest is.

Certain gestures, most obviously pointing, clearly relate to controlling the region objects and setting the region of interest. This suggests the region object as a concept to help us interpret the actions that the instructor might make and how these could influence videography decisions. Often the focus is "controlled" by verbal and high-level semantic information. However, some more syntactic cues that could potentially be identified through gesture recognition include:

- writing (to create regions), or erasing (to destroy regions);
- pointing, or some variant (such as tapping on the board) to direct the focus;
- waving or combinations of pointing gestures to link regions on the board.

4.4 Processing and Assembling Regions

The focus, and other gestural information, provide cues as to what should be shown to the viewer and, thus, where their attention should be guided. The system has the freedom to choose the visual methods used to realize this communication. The region object concept provides a mechanism for not only suggesting what should be the focus, but also how it is brought to the focus.

At any instant, the audience sees (and the captured video records) a number of region objects in a particular state and spatial arrangement. While the most obvious way to produce a video is to simply recreate this same arrangement, this is not necessarily the only option. The system has flexibility in choosing which region objects to show, which version of them to show, how they are shown, and how they are arranged within the video screen. Similarly, the system has a choice in whether or not to include the instructor layer of the video at any instant.

Composition of shots, and the sequencing of shots, must be done with regard to cinematic convention. This is important to preserve the communicative potential, to avoid jarring the viewer, as well as to provide more pleasing results. For example, basic rules of editing suggest restrictions on what types of shots should follow other shots to avoid making a "jump cut" (a jarring change of viewpoint that can distract the viewer).

4.5 Composition Examples

By casting the problem of shot composition and selection as the processing and combination of interest regions, we not only open ourselves to a new range of effects and manipulations not typically considered with video, but we also gain insight into how the composition might be cast as a computational problem. To illustrate this, we take a specific example from a lecture, shown in Figure 2. In this example, the lecturer has drawn several figures, and formulas on the board. In this case, he is trying to draw the classes attention to a matrix, by pointing at it.

This example emphasizes some obvious reasons why translation is important for moving from the medium of lecture to the medium of video. In this particular classroom, the aspect ratio of the chalkboard is quite different than in video: choosing to show the entire chalkboard would require inefficient use of the video frame. While the gesture used in the lecture clearly suggests which region should be focused on, this gesture also obscures the region. Additionally, in the video frame, the emphasized region is small. While having the additional context of the other regions is useful on the chalkboard, the limited space of the video screen makes such context an expensive luxury as the focus is too small to be seen clearly. Finally, the layout of the regions of the board is an artifact of the progression of the lecture.

Figure 2: The image used for the composition example. On the left shows the video image at the instant of time in question. The center shows a slightly later time, allowing the region behind the instructor to be determined. The right shows the complete chalkboard, with region objects defined.

Figure 3: A medium shot created by expanding the instructor and region of interest to fill the frame.

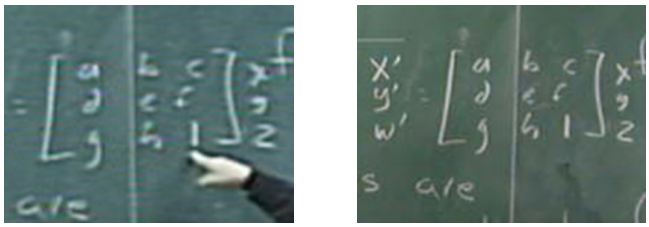

Figure 4: An extreme close-up shot (ECU) created by filling the frame with the region of interest from the original video image. On the right, a later frame from the still camera is used to create a clearer, unobstructed view.

Our goal is to compose a picture that provides a strong cue to the viewer as to what the focus is and makes it clearly visible. To do this, we can choose any of the available elements, process them in any desired way, and compose them in the frame as needed. The most obvious solution, merely reproducing the incoming video, relies on the cues of the lecture for attention direction. In film terms, this is a wide shot.

Other standard film shots can be cast in terms of the regions and focus. For example, a closer shot (arguably a medium shot) could be created by filling as much of the frame as possible with the two "participating objects," the instructor and the region of interest (see Figure 3). An extreme close-up shot, which in film terms focuses only on the detail of interest, could fill the frame with the focus object, as in Figure 4. This clearly is a mechanism for emphasizing the focus object. To improve this further, we could choose an unobstructed form of the focus object from the best possible view (the digital still camera), as seen in the right side of Figure 4. Because there is no motion involved, using this still is effective.

Through choices in what regions are shown, how these regions are processed, and how they are layed out in the frame, more composition opportunities are available to a Virtual Videography system. For example, to improve the above medium shot, we can composite the current frame with a future frame so that the instructor is made semi-transparent (Figure 5).

Figure 5: Compositing with a future frame allows the text to be read even with the instructor obscuring it.

Applying image processing operations to each region independently allows us to emphasize, or de-emphasize, regions as needed. Figure 6 shows various examples of applying image processing filters in this way to achieve the emphasis. Notice that in each case, we are able to emphasize the region of interest without obscuring

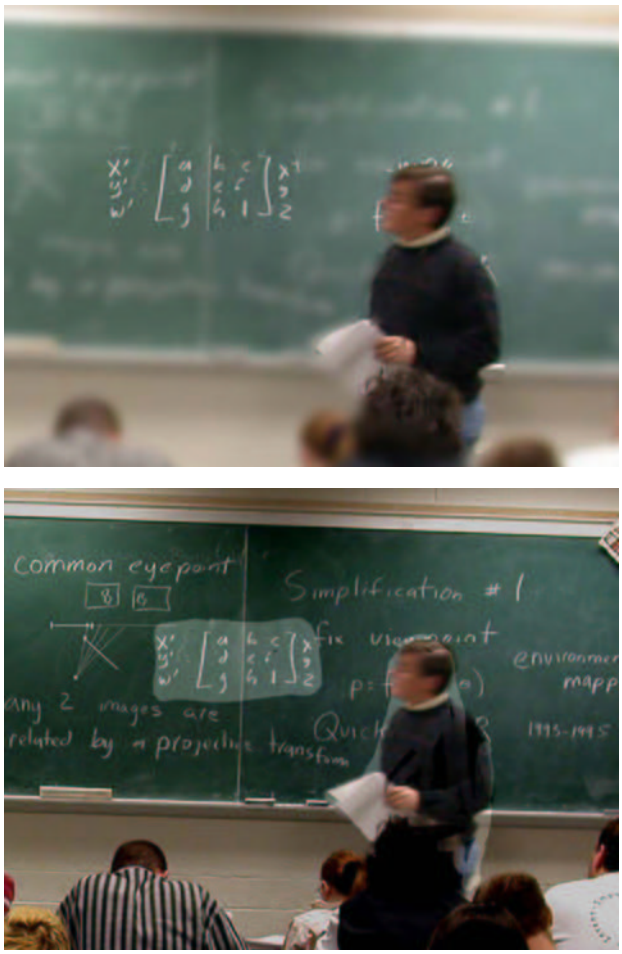

Figure 6: Image processing operations are applied to regions independently. This provides a "stylized" mechanism for guiding the viewers focus. In the upper image, non-focus objects are blurred to deemphasize them. The lower image emphasizes the focus object by brightening it.

it. Similarly, other graphical representations of a "pointer" can be used to guide the viewers attention, as shown in Figure 7. Notice that in these cases, the system need not know what is in the regions, only that they exist and that one of them is to be emphasized.

A final degree of flexibility is that we can arrange the region objects as desired to best fit on the video screen. While being able to select a set of region objects, and an appropriate arrangement, may require more information, there might be some cases where the low-level information suggests such an arrangement. For example, if a number of objects are gestured at to become the focus in rapid successor, we might arrange them in a list, as shown in Figure 8.

The art of film editing has evolved to provide many styles and techniques that permit its best practitioners to achieve a variety of effects. For classroom Virtual Videography, our goals are much less lofty as we do not aspire to the high art of editing, nor does it seem necessary for the domain. We do, however, need to consider basic rules of editing to avoid creating video that is distracting or boring. Therefore, a Virtual Videography system must follow some basic rules of film editing when making the final decisions about which shots to generate. The most obvious rule is that a variety of shots should be used in order to avoid boring the viewer. Along the same lines, no shot should be less than four seconds or longer than twenty seconds. These simple rules help dictate what shots the system will create at different moments in the final video.

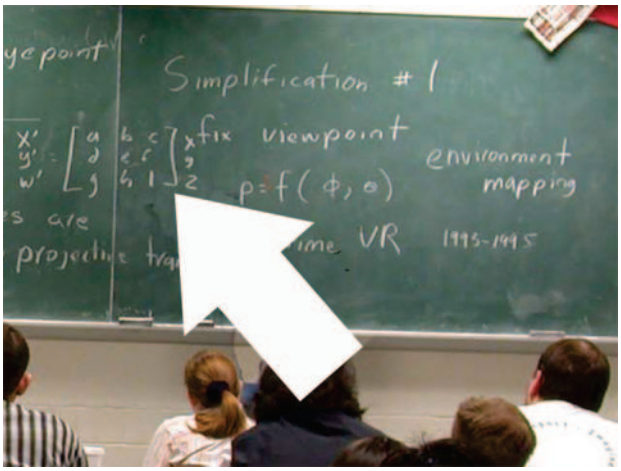

Figure 7: A graphic is used to guide the viewer's focus.

5 The Virtual Videography Project

The prospects of a Virtual Videography system are enticing: there is value in having video records of many types of events, especially if the video can be captured unobtrusively and produced automatically. While our focus has been on classroom lectures, there are many other types of domains that would benefit. For example, we have been considering meetings both of groups and interviews, design reviews, and question and answer sessions.

To successfully construct a Virtual Videography system, we must combine video capture techniques to obtain the source footage, computer vision to provide clues as to what should be shown, a degree of knowledge of cinematographic decision making to compose and select shots, and image synthesis methods to create the resulting decisions.

To date (February, 2002) we have not built a Virtual Videography system. We have not demonstrated that the daunting task of Virtual Videography can be solved. There may be reasons to doubt that it is possible. The limited information that would be available to a system from its vision components and limited techniques for synthesizing images may be insufficient for creating video with sufficient quality to hold a viewers attention.

Our work has provided a framework for Virtual Videography that suggests that we can make use of the limited syntactic information provided by computer vision for making certain types of cinematic decisions. The region object framework limits the demands on the vision system to the (seemingly reasonable) task of identifying related regions and to identify gestures that select regions as the focus. The framework also provides ways to use simple image processing and resampling to create a wide range of compositional effects. We must still develop algorithms for identifying regions, determining the focus, composing the images, and switching among shots. These pieces must be integrated into a system capable of handling the large amounts of data resulting from a lecture's worth (or a course's worth) of video. We must also evaluate the results of such a process at a large scale to see if the video has sufficient quality to be watchable and sufficient value to be useful.

Acknowledgements This work was supported in part by NSF grants CCR-9984506 and IIS-0097456, Micrososft Research, and equipment donations from IBM, NVidia and Intel.

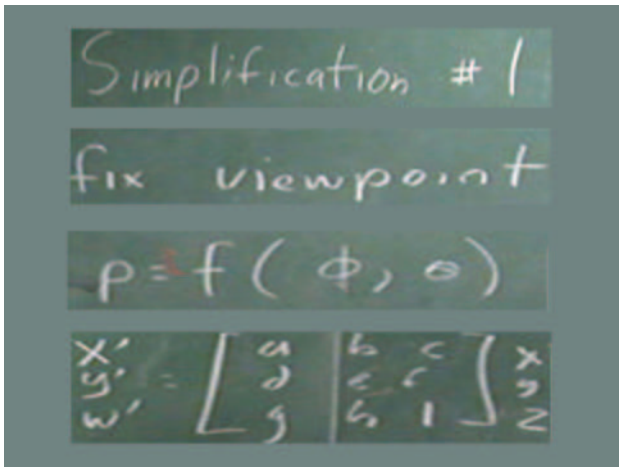

Figure 8: Region objects are re-arranged into a list.

References

[1] William Bares and James Lester. Intelligent multi-shot visualization interfaces for dynamic 3d worlds. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, pages 119–126, 1999.

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

[3] David Bordwell. *On the History of Film Style*. Harvard University Press, 1998.

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

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

[6] Foveal Systems, LLC. Auto auditorium. web page, 1999-2000. www.autoauditorium.com.

[7] William T. Freeman, David B. Anderson, Paul A. Beardsley, Chris N. Dodge, Michal Roth, Craig D. Weissman, William S. Yerazunis, Hiroshi Kage, Kazuo Kyuma, Yasunari Miyake, and Ken ichi Tanaka. Computer vision for interactive computer graphics. *IEEE Computer Graphics & Applications*, 18(3):42–53, May - June 1998. ISSN 0272-1716.

[8] Michael Gleicher and James Masanz. Towards virtual videography. In *Proceedings ACM Multimedia 2000*, November 2000.

[9] 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.

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

[11] 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.

[12] Michael J. Jones and James M. Rehg. Statistical color models with application to skin detection. Technical Report CRL 98/11, Cambridge Research Laboratory, December 1998.

[13] Shanon Ju, Michael Black, Scott Minneman, and Don Kimber. Summarization of video-taped presentations: Automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video Technology*, 1998.

[14] 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.

[15] Norman J. Medoff and Tom Tanquary. *Portable Video ENG and EFP*. Focal Press, fourth edition edition, 2002.

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

[17] University of Wisconsin Madison. eteach. http://eteach.engr.wisc.edu/newEteach/home.html, 1999.

[18] Yong Rui, Liwie He, Anoop Gupta, and Qioung Liu. Building an intelligent camera management system. In *Proceedings of MultiMedia*, 2001.

[19] Eric Saund. Image mosaicing and a diagrammatic user interface for an office whiteboard scanner. http://www.parc.xerox.com/spl/members/saund/zombieboard-public.html.

[20] Harpreet Sawhney and Serge Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, April 1996.

[21] Arno Schodl, Richard Szeliski, David Salesin, and Irfan Essa. Video textures. In *Proceedings of SIGGRAPH 00*, 2000.

[22] Tanveer Syeda-Mahmood and et al. Cuevideo: A system for cross-modal search and browse of video databases. In *Proceedings of Computer Vision and Pattern Recognition*, June 2000.

[23] Virtual Ink Corp. Mimio. Computer Hardware Product, 2000.

[24] Michael N. Wallick, Niels da Vitoria Lobo, and Mubarak Shah. Computer vision framework for analyzing computer and overhead projections from within video. *International Journal of Computers and Their Applications*, 8(2), June 2001.

[25] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *The IEEE Transactions on Image Processing Special Issue: Image Sequence Compression*, 3(5):625–638, September 1994.