WHY COMMON SENSE FOR VIDEO PRODUCTION?

Barbara Barry and Glorianna Davenport
{barbara, gid}@media.mit.edu
Media Laboratory, Massachusetts Institute of Technology

ABSTRACT

Video cameras are becoming cheap, small and ubiquitous. With advances in memory, cameras will increasingly be designed to be always ready, always recording. When cameras are always ready, how will videographers – professional and/or amateur -- decide what to shoot, when to shoot and how to index their video material to best support their communication requirements? In this paper, we describe an approach and early experiments that use a commonsense database and reasoning techniques to support a partnership between the camera and videographer during video capture. We describe a new paradigm for producing commonsense video metadata and describe how it can have a positive impact on video content capture, representation, and presentation.

Keywords: Documentary videography, video content indexing and retrieval, commonsense knowledge, video production

1. INTRODUCTION

For the last three decades, the size and price of the consumer video camera has steadily decreased while the quality of the recording has increased. Today, "webcams" and pocket video cameras are firmly entrenched, and desktop video editing packages are becoming ubiquitous. Meanwhile, tiny cameras are now available which can support the design of wearable outfits where the consumer can record anything and everything without necessarily making any conscious decisions about when to turn on, turn off or where to point the camera.

The challenge of portable, persistent camera is one of direction. How can appropriate images be captured that communicate a story about what was observed? And, what available representation can be drawn on that can assist the acts of selection and sequencing?

We believe that a persistent camcorder requires an active partnership between the consumer and the device. This partnership might be well served by creating a system that could cue the videographer and/or the camera about what and when to record, what to look at, and how to frame the image for an aesthetically pleasing edit? The question at hand is therefore how to architect a system that can provide cues to both partners (camera and human) to help them capture shots such that they will combine into strong sequences? Can commonsense reasoning help solve these issues?

2. WHY COMMON SENSE?

In the 1920's Kodak invented the Brownie camera. In the 1930's and continuing into the 1950's Kodak enhanced their marketing with basic tutorial booklets whose subject was how to shoot a good still picture. This marketing effort affirmed that the activity of photography required more than simply snapping the picture; it required that the consumer decide what to capture and when.

Video differs from still photography in that it captures a series of images over time. When played back, this serial representation of the world conveys motion, action and story. Sergei Eisenstein, the influential Soviet filmmaker and theorist of the 1920's, described filmmaking as the dual process of skillfully capturing the immutable shot and creatively combining the shots by juxtaposition into sequences conveying meaning, an editing technique which he called montage [4]. Normally, a good shot is stable and clearly conveys the intention of the author relative and an action, character or place. Often it involves a unique visual perspective, an attribute that should relate to other shots in the sequence. A sequence is made up of one or more shots that together communicate a complete idea, change in circumstance, or action.

A scaffold for the filmmaking process should provide two types of suggestions: (1) suggestions that relate to the capture of each individual shot, and (2) suggestions that emerge from an understanding of story and its relationship to particular shots.

In "Society of Mind," Marvin Minsky introduces the idea of common sense by showing us the many things a child must learn before being able to build with building blocks. "Common sense is not a simple thing. Instead it is an immense society of hard-earned practical ideas – of multitudes of life-learned rules and exceptions, dispositions and tendencies, balances and checks" [6]. Like learning to build with blocks, learning to capture a video sequence that conveys the desired meaning and emotion requires that the videographer learn many lessons.

Most importantly, videographers learn through experience. The more they shoot and edit, the more they are able to control their camera to capture the interesting detail of a situation in a way that admits discontinuity in time. Shooting draws on our ability to build an understanding of our perception of what is taking place, as well as an understanding of what is contained in the frame. How do the borders of what is seen through the camera lens constrain and focus the representation of the world? How can that partial representation be joined to another partial representation? And, does the viewer reconstruct meaning from the aesthetics of this joining?

Videographers refine their craft through the practice of editing and sequencing their shots. Through this delayed feedback process—filtered through an idiosyncratic knowledge of cinematic form and the recollection of previously captured shots in the scene – a good videographer ultimately learns to augment and adjust a model of the final sequence while framing a continuing set of shots.

As part of the skill, videographers soon gain mastery of "two-eyed" shooting, where one eye watches the image in the viewfinder while the other scans the entire scene to anticipate where significant action will occur. This skill is critical to moving the camera correctly.

If we could build the knowledge of the videographer into a system that could provide suggestions in real time to the videographer who is using a persistant camera, it may be possible for the machine to become a partner, helping the videographer reason about what will make a good sequence.

3. WHY THIN DESCRIPTIVE DATA IS NOT ENOUGH

Metadata is descriptive information assigned to multimedia that enables computational searching, associating or understanding of content. Increasingly, with today's cameras, certain metadata – such as time, date, GPS coordinates, compass information -- can be captured automatically. Additional metadata can be extracted through sound and image processing and linked to video segments. Still other metadata – such as keywords about who what, when, where -- can be manually attached to video shots following shooting.

Significant research has been done on the potential of using this metadata to retrieve and even sequence video [2,3]. Researchers have also shown that layers of metadata can be built up over shots to more precisely pinpoint where certain events occur in the content or format of a shot [1,5].

Most software developed to help support documentary filmmakers, whether novice or expert, has focused on providing a post-production environments for annotation and video editing. Alas, all to often, we discover a fatal flaw, a lack of detail, or story development is discovered after shooting has stopped and the characters have scattered.

Annotation captured automatically or attached after the recording moment, can do little to insure that the videographer captures a compelling rendition of an action which is unfolding. Moreover, most annotation is "thin;" it does not come with a rich compendium of practical ideas, life learned rules, exceptions, dispositions that we use every day in our reasoning about the world.

4. THE PROMISE OF COMMONSENSE

Common sense is the collection of knowledge and methods of reasoning we use to make sense of the everyday world. Although we make use of common sense during our daily life, in conversation, in actions, activities, this knowledge is rarely made explicit. For example, if I tell you that I am drinking a glass of water you know many implicit facts about this activity – there will be less water in the glass when I am finished, at some point the glass was filled with water, water is a liquid that can be spilled.

In Artificial Intelligence research, commonsense reasoning promises a means for computers have more human-like intelligence. Using large collections of commonsense knowledge and various methods of reasoning, computers will have the capacity to understand and make decisions about everyday situations. In a recent example, Erik Mueller [7] built a calendar application called *SensiCal* that uses commonsense from his Thought Treasure database. The purpose of SensiCal is to combine scheduling with commonsense reasoning to alert the user about mistakes in scheduling and fill in information that is not explicit in the calendar entry. For example, if the user tries to schedule a dinner date at 2am the calendar would question the entry. It knows dinner is usually in the evening hours of the day, and that restaurants may stop serving food before the wee hours in the morning. This application demonstrates how commonsense is useful for computer applications used in human contexts.

What if we use commonsense knowledge to reason and direct us in our video capture? Can it help us get better (1) shots, and (2) sequences? Can it create a richer environment for a partnership between videographer and a persistent camera?

We are using Openmind Commonsense (OMCS) as the main resource for dynamically accessing commonsense metadata for video content. OMCS is the first commonsense knowledge database amassed through public contribution on the WWW [9]. Users can go to the site and enter facts and stories (http://openmind.media.mit.edu), which contribute to a quickly growing, public database of knowledge. During the video capture process this knowledge can be dynamically accessed for use as metadata. The vidoegrapher can also be a contributor to OMCS. The strength of OMCS is the reciprocity, the video capture is informed by commonsense knowledge and the experience of shooting an event can inspire the videographer to contribute to the OMCS repository.

5. COMMONSENSE FOR DOCUMENTING LIFE

Video documentary making involves a dynamic process of collecting images, predicting, selecting and connecting video clips to communicate an idea or story to an audience. Computers can be used for all of these tasks when applied in a limited domain, one so limited that every possible decision must be described to the system in advance. The real world in front of the camera lens is diverse and surprising. Cameras might be smarter if they could use commonsense to understand the world in front of the lens.

There are two types of commonsense knowledge that can impact the shooting process – *formal sense* and *subject sense*. We can use *formal sense* in gathering the specific kind of commonsense used by experienced videographers. Examples of this formal knowledge are:

- *When shooting a conversation, record dialogue and reaction shots from each member of the conversation*
- *Take close ups of intricate actions to communicate the activity to the audience.*

- *Let the camera roll a bit beyond where you think the action stops in case you did not anticipate the ending correctly.*
- *When someone is walking, take a shot of where they came from and where they are going.*

This knowledge can guide an inexperienced videographer or remind a seasoned one about techniques that will improve a sequence. This kind of commonsense can be acquired from videographers' stories of successes and failures, and therefore can be invested in a commonsense database.

Subject sense is knowledge about the people, places, events, and situations the camera will encounter and record. Here are a few examples:

- *Running is faster than walking.*
- *An effect of running is movement.*
- *People run when they are being chased.*
- *People run to exercise.*
- *People do not run and eat at the same time.*

For a camera to understand the significant, unusual and complex story significance of a shot, it can use commonsense to reason about how a shot contributes to a story thread represented by a sequence of individual shots. In documenting life, each captured shot is a potential starting point of another story thread. If the computer can predict possible stories that a newly acquired shot could incite, as well as understand the strong or weak relationships between the shot and previously acquired material, it can make powerful suggestions that can help the videographer construct a landscape of content that can later yield coherent and creative scenes in the editing phase.

In the following sections we present scenarios to express our vision of the partnership that could occur between the videographer and the camera during shooting enabled by commonsense knowledge.

6. MARATHON EXAMPLE

Your goal is to shoot a marathon. The truth is, you've always wanted to run in one and perhaps capturing this marathon will give you a bit of inspiration. When you think of a marathon chances are that images come to mind such as runners, sweat, crowds cheering, and water cups passed to tired athletes. Our minds conjure a story at the word, "marathon." You've never been to one before so perhaps the pictures you have are from a television show or a story a friend told you at a party last year.

You go to the marathon. You shoot some video. You come home and watch it. It does not really give you any sense of the experience. You were hoping since you watched practically the entire event through the LCD or viewfinder that the video images would reveal something about the event that you did not see during it. There are 3 hours of footage, most of people running and people watching people run. You shot the event but what are the qualities of a marathon that make an intriguing story? You put away the footage and never watch it again.

Video has the power to amplify our impressions of an event through putting a magnifier to the details of the drama, to what we find intriguing and compelling. Like our memories, it is not merely a tape recorder.

Let's go the marathon again. This time the camera is your partner. You tell it your goal. You want to shoot a marathon. It knows a lot about marathons. Here are a few examples of what it knows:

- *Runners often eat pasta the day before the race.*
- *At the end of the race the runners are exhausted.*
- *Not every runner crosses the finish line.*
- *The starting line is where the race begins.*
- *Runners pick up their numbers to wear before the race begins.*
- *People cheer to encourage the runners.*

The camera can also retrieve a more constrained script that could be used as a simple, temporal shot list [8].

- *The runners line up at the starting line.*
- *The runners start at the sound of a gun.*
- *The runners run the length of the marathon.*
- *One runner wins by crossing the finish line first.*
- *The crowd cheers.*

This subject sense knowledge about marathons can be enhanced by formal sense knowledge to create shot suggestions encouraging the videographer to capture a diverse or specific cannon of clips that could later be easily assembled into a story. When a subject sense suggestion is selected as a shot that the videographer does take, formal sense knowledge could be incorporated to generate a shot suggestion. This is accomplished through commonsense inference and reasoning by the system. Here are two examples:

Example 1:

- *Subject sense: Runners start at the sound of the gun.*
- *Formal sense: Shoot events that catalyze other actions in extreme close up then get a shot of the action triggered.*
- ➤ *Initial shot suggestion: Shoot a close up of the gun being fired.*
- ➤ *Following shot suggestion: Shoot the runner's crossing the starting line.*

Example 2:

- *Subject sense: At the end of the race the runners are exhausted.*
- *Formal sense: Shoot actions to show internal states.*
- ► *Shot suggestion: Shoot runners slowing down after the finish line.*
- ► *Shot suggestion: Show runners not able to walk without assistance.*
- ► *Shot suggestion: Show runners trying to catch their breath after just crossing the finish line.*

Ideally, the vidoegrapher should be able to choose to view the commonsense knowledge, the shot suggestions or both in the camcorder viewfinder intermittently during the event. The commonsense knowledge should be associated with the shot for possible later use in the editing process of production. We are not suggesting that the camera absolutely direct the shooting process, but that it becomes a creative partner in directed shooting to better support the later composition of shot sequences to convey the marathon.

7. FIRST EXPERIMENTS

We developed a desktop application in which videographers can upload video clips and annotate them with simple English sentences. The system queries the OMCS database with the English description and returns relevant commonsense knowledge and associates it with the video clip, thereby expanding the context of the clip. This application will be integrated into the capture process by installing it on a portable, computationally-enabled, wireless camera.

We are working with Singh on a configuration that can give videographers a way to dynamically input and access their own collection of commonsense knowledge in OMCS as an addition to their access to the entire database of knowledge contributed by thousands of anonymous authors. This will allow us to grow collections of commonsense knowledge specific to videography.

8. CONCLUSION

When the camera becomes a partner to the videographer, able to understand, organize and make suggestions about video shots and sequences, the documentary making process can become more closely integrated with life. Commonsense can help in creating this partnership by providing the camera system with the ability to reason about life situations we choose to record.

Bringing heightened awareness of the content landscape to both the filmmaker and the camera during the shooting/production process not only can serve to close gaps in content resulting in higher success in editing story sequences but also can illuminates alternative story ideas to encourage creative documentary videography.

9. ACKNOWLEDGEMENTS

This research is supported in part by the MIT Media Lab Digital Life and Information Organized consortia as well as by the Motorola Corporation. We would like to thank Push Singh for his many contributions to our thinking and to this paper. Tara Rosenberger Shankar helped to refine this paper with her valuable editing suggestions.

10. REFERENCE

[1] Davenport, G. and T. Smith (1991). Cinematic Primitives for Multimedia. *IEEE Computer Graphics and Applications. 11*(4). p 67-74.

[2] Davenport, G. and Murtaugh, M. (1997) Automist Storyteller Systems and the Shifting Sands of Story. *IBM Systems Journal. 36*(3), p 446-456.

[3] Davis, M. (1994). Media streams: Representing video for retrieval and repurposing. *In Proceedings of the ACM Multimedia Conference 1994*, San Francisco, ACM, 1994. p. 478-479.

[4] Eisenstein, S. (1949). *Film form: essays in film theory*. New York, Harcourt and Brace.

[5] Kankanhalli, M. and Chua, T. (2000). Video Modeling Using Strata-based Annotation. *IEEE Multimedia. 7*(1) p. 68-74.

[6] Minksy, M. (1985). *Society of Mind.* New York: Simon and Schuster.

[7] Mueller, E. (2000). A Calendar with Common sense. *In Proceedings of International Conference on Intelligent User Interfaces*, New Orleans, ACM, 2000. p 198-210.

[8] Schank, R. and Abelson, R. (1977) *Scripts, Plans, Goals, and Understanding.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

[9] Singh, P. (2002). The Public Acquisition of Commonsense Knowledge. *In Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*. Palo Alto, CA, AAAI, 2002.