DOCUMENTING LIFE: VIDEOGRAPHY AND COMMON SENSE

Barbara Barry, Glorianna Davenport

MIT Media Lab, 20 Ames St., Cambridge, MA 02139 USA
{barbara, gid}@media.mit.edu

ABSTRACT

This paper introduces a model for producing common sense metadata during video capture and describes how this technique can have a positive impact on content capture, representation, and presentation. Metatada entered into the system at the moment of capture is used to generate suggestions designed to help the videographer decide what to shoot, how to compose a shot and how to index their video material to best support their communication requirements. An approach and first experiments using a common sense database and reasoning techniques to support a partnership between the camera and videographer during video capture are presented.

1. INTRODUCTION

Video cameras are becoming cheap, small and always ready, always recording. When cameras are always ready, how will documentary 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? The challenge of portable, persistent camera is one of direction. How can appropriate images be captured that communicate a story about what was observed? What available representation can be drawn on that can assist the acts of selection and sequencing?

A persistent camcorder requires an active partnership between the videographer 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 problem at hand is therefore how to architect a system that can provide cues to both partners (camera and human) to help them capture shots that will combine into strong sequences. Common sense knowledge and reasoning can support camera understanding of event relationships within the documentary subject and the fundamentals of effective camera work.

2. WHY COMMON SENSE?

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 conversations, actions and activities, this knowledge is rarely made explicit. For example, upon observing someone drinking a glass of water there are many facts we know about the event – there will be less water in the glass when they are finished, at some point the glass was filled with water, water is a liquid that can be spilled.

In "Society of Mind," Marvin Minsky describes the complex common sense 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" [7]. Like learning to build with blocks, learning to capture a video sequence that conveys the desired meaning requires that the videographer learn many lessons. Most importantly, videographers learn through experience. Shooting draws on the ability to build an understanding of the perception of what is taking place, as well as an understanding of what is contained in the frame.

A scaffold for the filmmaking process should provide the following two types of suggestions: (1) suggestions that relate to the capture and framing of each individual shot, and (2) suggestions that emerge from an understanding of the function of a particular shot in a sequence of events.

3. FILMMAKING 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 and how to capture an image.

Video differs from still photography in that it captures a series of images over time, video clips or shots that are often concatenated into sequences. A sequence is one or more shots that together communicate a complete idea, change in circumstance, or action [1]. Still photography and videography share a high-level goal – a good shot is stable and clearly conveys the intention of the author relative to an action, character or place. In videography, each shot involves a unique visual perspective, an attribute that enables continuity by relating to other shots in a sequence and to the global logic of the documentary [8]. Videographers learn the technique 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 in accordance with the flow of events unfolding in front of the camera. A good videographer ultimately learns to augment and adjust a model of the final sequence while framing a continuing set of shots. A good

videographer creates a collection of clips that best supports the documentary story.

4. COMMON SENSE IN ARTIFICIAL INTELLIGENCE

In Artificial Intelligence research, common sense reasoning promises a means for computers have more human-like intelligence. Using large collections of common sense 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 built a calendar application called *SensiCal* that uses common sense from his Thought Treasure database [8]. The purpose of SensiCal is to combine scheduling with common sense reasoning to alert the user about mistakes 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 will question the entry. It knows dinner is usually in the evening hours of the day, and that restaurants may stop serving food before the early hours of the morning. The system must understand the details of a dinner scenario in order to predict events, identify contradictions and track continuity over events. This application demonstrates how story understanding using common sense is useful for computer applications used in human contexts.

Current approaches to story understanding by a camera are concerned primarily recognizing series of events in very limited domains, most successfully in analysis of sporting events such as soccer or highly scriptable domains such as cooking shows [5,10]. These systems inherit a problem from all symbolic story understanding research. They are successful in extremely limited domains and demand the engineer hand code into the system representations for all known story events, details and outcomes. Camera understating of the dynamic world in front of the documentary maker's lens requires broad, deep and flexible knowledge. Recent advances in commonsense reasoning and information retrieval support the understanding of broader story domains approximating the real world [8,11]. The need for large common sense databases has been sited as absolutely necessary for advances in story understanding [6].

Openmind Commonsense (OMCS) (http://openmind.media.mit.edu) is the second largest common sense database, after Cyc [13,4]. OMCS is the first common sense knowledge database amassed through public contribution on the WWW [13]. Users contribute common sense information such as explanations of events, relationships between words and the functions of objects. The advantage of OMCS over other common sense databases for use in documentary videography is three fold. First, OMCS has the potential for reciprocity. The common sense knowledge supports the activity of video capture and the experience of shooting an event can motivate the videographer to contribute to the OMCS repository. This is enabled by the representation of knowledge in natural language instead of a special formal representation such as Cyc-l which much be learned by each contributor. Second, the OMCS knowledge is free and available for download. Third and most importantly, OMCS has significantly more knowledge related to sequencing of events than any other common sense database. The camera knowledge about event relationship is critical for successful reasoning about sequences of events during capture.

5. IMMEDIATE AND HIGHLY DESCRIPTIVE METADATA

In order to give immediate feedback to the videographer during the shooting process metadata must be aggregated at the time of capture. *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, and 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 associated with video shots after shooting has been completed.

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 [1,3,5].

Most software developed to help support documentary filmmakers, whether novice or expert, has focused on providing 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 long 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.

6. COMMON SENSE 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. A camera must reason about the filmmaking process and the content being captured to be a useful partner. Therefore, there are two types of common sense knowledge that can impact the shooting process – *formal sense* and *subject sense*. *Formal sense* is defined in this paper as the common sense knowledge about filmmaking used by all novice and seasoned videographers. In contrast, expert filmmaking knowledge would be used to help novice videographers emulate the style of a particular filmmaker or genre of documentary filmmaking. Although useful, the focus of this paper is formal sense, general knowledge about shooting competently to support a broad scope of documentary styles and subjects. Examples of formal sense 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.*
- *When someone is walking, take a shot of his or her origin and destination.*

This knowledge can guide an inexperienced videographer or remind a seasoned one about ideas that will improve a sequence.

This kind of common sense can be acquired from videographer reflections on shooting technique and invested in database of formal sense knowledge.

Subject sense is knowledge about the people, places, events, and situations the camera will encounter and record. It is the common sense knowledge necessary for understanding the potential story of the documentary subject. Here are a few examples:

- *Running is faster than walking.*
- *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 common sense to reason about how a shot contributes to a story thread or sequence of individual shots. When documenting life, each captured shot is a potential starting point of an additional 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.

The following scenario expresses a vision of the partnership between the videographer and the camera enabled by common sense knowledge feedback during shooting.

7. MARATHON EXAMPLE

Your goal is to shoot a marathon. When you think of the race images come to mind such as runners, sweat, crowds cheering, and water cups passed to tired athletes. Your mind conjures a story at the single word, "marathon." On marathon day you shoot some video by watching almost the entire event through the LCD viewfinder. The captured footage is three hours of crowds, people running and people watching people run. The footage does not communicate your experience of viewing the marathon, the story of the day or your interest in the subject. 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 by 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 send you to the marathon again. This time the camera is your partner. The camera does not absolutely direct the shooting process but becomes your creative partner in directed shooting to better support the later composition of shot sequences conveying the story of the marathon. Here are a few examples of what your camera knows about marathons:

- *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.*
- *People cheer to encourage the runners.*

Your camera can also retrieve a more constrained script that could be used as a simple, temporal shot list [10].

1. *The runners line up at the starting line.*
2. *The runners start at the sound of the gun.*
3. *The runners run the length of the marathon*
4. *One runner wins by crossing the finish line first.*
5. *The crowd cheers.*

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. This is accomplished through common sense inference and reasoning by the system. Here is an example:

- *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 firing.*
- *Following shot suggestion: Shoot the runner's crossing the starting line.*

Ideally, the videographer has access the common sense knowledge and shot suggestions intermittently during the event through a visual or audio channel. The common sense knowledge accessed during shooting is permanently associated with the shot for possible later use in the editing process of production.

8. FIRST EXPERIMENTS

Currently a wearable system functions as the computationally enabled video camera. It is comprised of the following elements: a 650MHz sub-notebook laptop with 10Gb hard drive, a 640x480 color heads up display, a shoulder mounted 640x480 firewire camera with 4mm focal length, a chording keyboard, rechargeable lithium battery supplies and an earbud speaker. The software is a java application enabling video capture; annotation of video clips with user input metadata descriptions; and, association of OCMS knowledge, which resides in an XML database. The applications overlaid on the observed scene through the heads up display.

Immediately following capture videographer annotates a video clip with an English sentence or phrase using the chording keyboard. The sentence is parsed and submitted as a query to a subset of the OMCS database. Relevant common sense knowledge is returned and displayed to the videographer who can use any result as a shot suggestion and can subsequently choose to permanently associate results with the video clip. This creates a fail-soft application in which the user is free to reject or accept suggestions and permanent annotations. Presently, five categories of OMCS data are used for subject sense feedback during the shooting process and as annotations for captured video. The categories are "the first thing you do", "the last thing you do", "cause and effect", "uses and functions" and "enter a fact". The first three categories supply relations between events. The fourth supplies relations between objects and functions. The last supplies free form knowledge. In the first shooting experiments it was observed that common sense feedback

Figure 1: Interface showing a captured clip, OMCS query results and video metadata.

delivered during the annotation of one clip influences the decision making process for capturing subsequent shots. This is particularly true for "the first thing you do" and "the last thing you do categories" as they suggest possible shots needed for a complete sequence. Number of results for each OMCS query varied greatly from one to over one hundred assertions illuminating the need for sorting and ranking algorithms.

9. FUTURE WORK

The first experiments provide a solid base and challenges. Increased reasoning capabilities by the camera such as statistical ranking of results and automatic generation of scripts from OMCS assertions are necessary. In addition, the subject and formal sense suggestions can by synthesized into a single shot suggestion to be accepted or rejected by the videographer. The next experiments using the system will be documenting a marathon, creating video sequences and comparing results to sequences shot without common sense feedback and immediate annotation during the shooting process.

10. 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. Common sense 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/capture process not only can serve to close gaps in content resulting in higher success in editing story sequences but can also illuminate alternative story ideas to encourage creative documentary videography.

11. ACKNOWLEDGEMENTS

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

12. REFERENCES

[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. (1995). Media Streams: Representing Video for Retrieval and Repurposing. Cambridge, Massachusetts, Massachusetts Institute of Technology.

[4] Guha, R., & Lenat, D. (1990). Cyc: A midterm report. *AI Magazine,* 11(3), 32-59.

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

[6] McCarthy, John, Minsky, Marvin, Sloman, Aaron, Gong, Leiguang, Lau, Tessa, Morgenstern, Leora, Mueller, Erik T., Riecken, Doug, Singh, Moninder, & Singh, Push (2002). An architecture of diversity for commonsense reasoning. *IBM Systems Journal*, *41*(3), 530-539.

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

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

[9] Nichols, B. (1991). *Representing reality: issues and concepts in documentary.* Bloomington: Indiana University Press.

[10] Pinhanez , C. & Bobick, A. (1996). Approximate World Models: Incorporating Qualitative and Linguistic Information into Vision Systems, Proc. of AAAI'96, Portland, Oregon, pp. 1116-1123, August 1996.

[11] Riloff, E. (1999). Information Extraction as a Stepping Stone toward Story Understanding, In Understanding Language Understanding, Ashwin Ram and Kenneth Moorman, eds., Cambridge: The MIT Press.

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

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