Sharing video memory: goals, strategies, and technology

Glorianna Davenport
Principal Research Associate
Media Fabrics Group
MIT Media Lab

Abstract: In this short paper, I explore video as a tool for recording and presenting our perception of reality.

Introduction:

What use is the video recorder? It is another eye and ear attached to a synthetic sense-memory that creates a time-based image record of some permanence. The act of capturing images with a video recorder allows us to explore our perception of reality as it unfolds; and in its re-viewing, video can support and extend our memory of past events. Video is often used to reinforce observation, reflection and communication in the workplace. It also provides many of us with a magnificent personal hobby.

What is required of us if we are to observe the world in a meaningful way? We must be at the right place at the right time; we must be filled with the right intent; we must comprehend what is going on in front of us and be able to predict how the action, at different temporal scales, might evolve into the future; and, we must feel confident in our partnership with the technology.

In the rest of this paper, I make the case that while the opportunity to observe and capture our perception of reality has never been easier or more accessible, the act of filming requires us to acknowledge our intentions about its making and use. Intention guides us to a plausible outcome; by that, I mean that the observations we make are narrative in nature, and the construction of a plausible narrative requires many conscious choices. These include: the choice of technology, how the technology is deployed, who controls the recording, how much freedom is there to compose shots, who controls the editing, who controls the channel of distribution, and who chooses to view the final result.

1st problem: What shall I capture?

The first problem we encounter when we have a camera in hand is, “What shall I shoot?” While this problem occurs with all kinds of cameras, the problem is exacerbated with video, as I shall explain as we go along.

When we first hold a camera it is easy to act impulsively. However, one can only wave a cell phone around and depress the record button so many times before we ask: what does what I have recorded mean? Who would like to see it? Is it

any good? What shall I shoot next? It only takes a moment to capture a still frame but that moment must carry meaning to the beholder. Capturing a video requires more time but is no less affected by the issue of meaning. The meaning can be more or less depending on the subject, the composition, the lighting, the sequence of shots etc. Decisions around these elements require us to invest concentrated time. Whenever we invest time for more than a few short periods of random fun, we are likely to ask: why? And if we can express the "why" in a satisfactory way, we might further ask: how can we do this better? This requires more time.

Taking pictures is a form of experimentation. We take a picture and then view the result. Does the artifact meet our expectation? At first this expectation is more like hope, tightly coupled to our emotional perception of the event and of our own situation in it. As we make more experiments, we think more about what we would like the results to communicate; our expectation tells us that our interpretation of the situation, our position, framing, lighting, and other momentary conditions of a situation all play a role in creating a "successful" final artifact. This experience feeds back into our expectation as we prepare to capture another image, sequence or event. In this way, the act of decision making frames and is framed by our intention.

Intention stirs the imagination. It is no use having a camera in our pocket as we walk down the street if nothing stirs our desire to use it. Time passes. Reality emerges and is transformed. It is no use remembering later and wishing we had captured a particular event in all of its momentary temporal splendor. Cameras capture moments; we have to record proactively. This requires taking the decision to capture something. Shall I video my child's first birthday party or a wedding? The process of constructing a Lego object or process building a house? A portrait of a friend cooking? A workshop with its interactions and creative processes? A marathon? Or shall I just carry a camera with me all the time and make a movie about nothing in particular, as Richard Leacock did in "Les oeufs a la coque?" Our lives are filled with events whose process and outcome are important to our future becoming. Can video augment the way in which we evolve an understanding? What do we want the image or the video to convey? Will we share the recorded moments with other family members and friends, with a professional colleague, or with the general public? Knowing some of this in advance allows us to better prepare. The deeper we probe, the more clearly we understand that to capture anything in a meaningful way requires intent.

Without intent it is difficult to know what to shoot, when to take a shot, from what angle, with what composition? When we compare video with still pictures, we come upon additional issues associated with time and motion. While the still photographer is capturing moments, the videographer is generally engaged in understanding and capturing an emerging phenomenon as it unfolds over time. The best documentary filmmakers learn to keep both eyes open, one looking

through the framed scene in the viewfinder and the other eye surveying the larger scene. It is vital to maintain a good understanding of what is going on, to anticipate what will happen next, and to decide where the camera should be pointed. Generally, the concentration required to understand and compose a scene as it emerges – as well as understanding what might be interesting if captured into a synthetic memory, and why -- takes our whole attention over some period of time

Of course, when we first hold a video camera our starting point may well be free-form experimentation. What is the camera capable of? What does this kind of motion communicate in playback? What strikes our fancy as we walk down the street? However, sooner or later we confront our own demons, those that want us to capture something relevant to our lives, something that will stir our imagination when we watch it in the future, something that we can share with others. At this moment, we become aware of story potential, a speculation into how the story we are witnessing might evolve. As we begin to direct our camera, we learn that we can turn the camera on and off, change position, and turn it on again to capture the next development. At this moment, we begin to understand sequence in a rich way.

We all know what a sequence is. It is the essence of cinema. We have all experienced sequences in dramatic films and television shows. The sequence is a fundamental unit of constructed meaning, a formal entity that juxtaposes two or more shots, compresses time, focuses attention on certain details within a scene, and provides the audience with a story unit. A scene is built up from one or more sequences. But how do we capture a sequence as the real-life action we are interested in is evolving? We have no script. We have only a sense of what is happening and what might happen and how what happens at this moment in time might relate to our larger story. These aspects frame our mental model which subsequently informs our actions as we find the best position for the camera and microphone, decide when to turn the camera on in time and turn it off, consider what we have shot, and reposition ourselves for what happens next.

The joy of capturing the sequence is the heart of hand-held documentary videography. Each time we record a sequence with documentary intent, it is an experiment. With each sequence, we draw on and contribute to our sense of what works and what doesn't, what is worth focusing on and what is not.

Patience, strategy, mistakes, recovery: a case in point

Recently my sister-in-law asked me if I would video her son's wedding. I agreed but made it clear that I would be filming throughout the wedding day, not just at the service itself. I knew the mother of the groom better than the groom himself. She is very talkative, enjoys life and was likely to be a good subject. I met the bride and her mother for the first time the night before the wedding as we arrived late at the rehearsal dinner.

The pre-nuptial events of the morning are nicely divided into potential sequences by character and activity. With my husband, I track down the various parties. At 9AM we meet the groom on the golf course. He has slept through his alarm clock's wake-up call and arrives late. I attempt to capture the ambience of his last golf game as a bachelor. Capturing any sequence her proved very difficult. How to use the scale of the golf course? How to be in the right spot to capture the interactions of the boys? Was that a good golf swing? The footage that I capture in an attempt to get to the heart of the thing will be greatly reduced in editing.

We lunch in the hotel with the older generation of family and close friends on the groom's side. My sister-in-law finally catches up with us and immediately spills the beans about the choreography of the procession. Each of the principals will be escorting a dog down the steps of an amphitheater! The delivery of this piece of news and subsequent reactions becomes the natural focus of the sequence. It also serves to inform my later shooting of the service.

Alas, I lost a fun scene after lunch of Mom getting dressed in my room. It was a typical but annoying mistake: I thought the camera was recording but it wasn't.

On the way up the mountain, we stopped to catch up with the bride. The bustle in this small cabin makes shooting a joy. There are wonderful moments of problem-solving: as the bride's mother helps a bridesmaid into her dress, a friend of the bride drops by to visit. Of course, mistakes are made: I do not have the camera turned on when the bride takes out her stash of toothbrushes and hands them around. Oh well… in the greater scheme of the day, not too important.

Once on the mountain, we wait around for almost an hour. The professional photographer is busy getting shots; the boys and dogs are milling around; people are meeting each other. I get a few shots but mostly puzzle about how to position myself in the very large outdoor amphitheater for the procession. Finally the wedding party is in motion and people seat themselves. My placement - to the left mid-way up the steps allows good quality sound of music, feet on the path and the dogs panting and me easy access to the stage area after everyone comes down. One difficulty in the service is that the sound is very diffused and augmented by the constant stream of jets that fly directly overhead. I video one just in case I need to make the point. I get a fun shot of the dogs in the front row just in the moment that the smallest of the dogs jumps into my sister-in-law's lap.

Getting a sequence out of the post-wedding festivities is trying. In the end, I resort to a favorite old trick: forget the toasts and speeches; get the dancing.

To summarize: in capturing an event, we use our mental model of the event to provide a means for breaking the event into scenes and sequences. Generally, scene breaks occur when we move to a new location or when a new set of

characters enters the location we are videoing. Sequences are chunks of activity that further the story. Sometimes scene and sequence overlap, sometimes there are a few sequences in a scene. Each sequence should not only further the overall story, but also communicate something new about a character or progress towards the character's goals. A different way of saying this is that when we capture actions, we are often also building portraits of the people involved.

In videoing almost anything, patience is a critical attribute: the best rarely happens first. Sometimes an activity goes on for a long time; we need only a little bit of the activity, but which is the salient bit? To use this wedding as an example: over the course of 12 hours, I captured only 52 minutes of video. I am able to limit the footage by thinking about sequences and anticipating events, making allowances for accidental discoveries and unexpected turns. A sequence allows me to focus, to turn the camera on and off, to shoot to edit. What I capture of the wedding is considerably less than the whole, and what remains after editing will be considerably less than that. However, I hope that the final film will give a sufficient impression of "what took place" and will give pleasure to those involved as they watch and remember.

2nd problem: what equipment should I use?

Filmmakers of all kinds enjoy talking about their gear. What is the make of your camera? How is it configured? What does it do exceptionally well? & etc. When purchasing a camera, many people will offer advice. However, to pick the right camera we need to understand what we are likely to use it for.

For the solo video-maker who wants to shoot hand-held, small is glorious. I try to avoid video cameras that look like guns or huge "professional" gear; these are not welcome in a crowd or in a small intimate surround. The modern palm-sized consumer cameras can be used in almost any circumstance. Of course, we still need to ask permission of our subjects but, overall, small cameras minimize the impact of recording on the dynamics of human activity. Even so, when we are shooting we often have to wait until people are doing things that are more important to them than noticing our camera.

The most significant difference between the professional and consumer cameras – besides their size -- is that the professional-grade cameras use 3 chips for imaging. Each chip contributes color information only in a certain range. This can result in better color and more resolution that the smaller single-chip consumer cameras provide, but there is a tradeoff in cost, size, visibility, and performance in low-light situations

Hand-held shooting requires particular concentration and synergy between the body, the eye and the camera. Shots need to be rock-steady stable. One achieves this by adopting the disciplined stance of a dancer. While newer video

cameras have flip-out monitors that allow us to see what we are shooting, this does not permit us to hold them more casually, something that beginners often assume. The intoxication of shooting often takes the beginner by surprise. This often results in a casual approach to holding the camera, framing and turning the camera off. Intention requires that we know our tools and our selves. Capturing a compelling sequence is not a matter of luck. One cannot create a sequence in the editing process when there are no cutting points or when the camera motion is so great that it will make any viewer seasick. Alas, for those first experiments! The tape that was so much fun to make will never be screened again, as not even a best friend has the patience to watch it.

So far I have focused on the camera, but what about sound? In the case of hand-held shooting, a good external microphone should always be used. In consumer video cameras, the built-in microphone is never satisfactory; they pick up a constant hum from the cassette drive and a variable hum from the zoom lens. Furthermore, when we are shooting single-person style, we need to get the microphone as close as possible to our subject. Remember the inverse-square law: “double the distance, a quarter the sound.” This means that issues of sound quality often limit the cameraperson’s choices of where she can position herself while shooting. The use of a directional microphone will further constrain how the filmmaker can move the camera: if we pan away from the speaker to a listener during an intense dialog, the sound level of the speaker will drop off precipitously. This creates a nightmare when editing. Richard Leacock has recommended one good solution: is to get someone to build you a double microphone which is highly directional on one channel and wide-angle on the other, plug it into the stereo input of the camera, and allow the editor to choose the best sound in any situation.

So much for the issues and equipment associated with one person recording a situation with a hand-held camera. Let us turn now to situations where one camera cannot sufficiently cover the action. For instance, if one tries to film a symphony orchestra with a single hand-held or fixed camera, one quickly discovers that such a recording is boring or worse, confusing. When we hear the tympani, we need to see the tympani up close not a shot of violins or the conductor. A symphony is continuous in time, and yet moments of interest occur at different parts of the stage at different times. The engagement of tympani is an unusual event in most symphonies and unusual events require our attention. We need multiple cameras if we are going to provide compelling visuals of the symphony in action. What strategies exist for multi-camera shooting? (1)

Television generally uses large studio cameras on stabilized platforms to continuously cover real-time events. Over the years, the television industry has invested millions of dollars in new equipment to cover real-time events efficiently and economically. Camera people who can follow a football when it is in play are handsomely rewarded, as are the commentators that allow the audience to follow the play-by-play action. By focusing on different parts of the action, multiple

cameras can provide a compelling representation of an event as it transpires over time. In the 1970's Bill Cosel made history by videoing and transmitting the first live performances of the Boston Pops Orchestra. Bill's method involved: first, marking up a score and assigning shots to different camera positions and, second, only hiring camera people who were able to sight-read music.

Increasingly, the scientific and professional communities have adopted video as a work tool. Many lectures are now captured in the back end of auditoriums and streamed over the internet. Other more difficult scenarios center on the imaging of very small or large-scale phenomena relative to the human scientist. In documenting a real-time surgical procedure for the record, a minimum of two cameras might be required: we need the micro perspective close up to the surgical incision and, to understand what is happening in the operating room, we need a wider, more distant shot. Again, if we want to "document" a surgical procedure for fictional entertainment, we need many more cameras – the true focus is the interplay of the actors -- but here we have moved into the unreal world of scripted and completely controlled activities, multiple takes and retakes, and extensive editing.

In my role as a Research Associate at MIT, students regularly propose marvelous new experiments. In the 1980s and early 90's many students came to me wanting to create a "head cam," a camera worn on a bicycle helmet. They argued that the captured video would show the world as a person sees it. My response was always fairly negative: too wobbly; since our eyes move within our head, a head-mounted camera wouldn't be focused where we were looking, etc. We see by means of difference; our eyes travel constantly and we compare what we just saw with a model of the space we are building our head. In video, our eyes travel around frames, looking out for differences between this frame and the last frame. It takes us two frames to mentally register a cut. I assume that too much change in sequential frames will ruin our ability to see exactly what has changed.

Six or seven years after I turned down one such student, Steve Mann -- one of the student "cyborgs" in Sandy Pentland's Vision and Modeling Group at the MIT Media Lab -- made a head cam. It was a stereo pair of cameras mounted on the sides of an eyeglass frame with small monitors in front of his eyes. (The images appeared upside down. I once asked Mann how long it took to get used to his gear. In the beginning it would take a few hours, he said, but after a while he was able to skip back and forth with ease. As he ran his hand along the bundle of cables running down his back, he said that he had the sense that this was his optic nerve.) A few weeks after that conversation, Steve wanted to discuss his shooting experiment. He had worn his odd-looking gear into a store and was forcefully confronted by the management: cameras were not allowed in their store, they said. Steve then pointed to the many security cameras mounted on their ceiling, ostensibly to capture shoplifters, and argued that he was merely doing the same thing: he thought of it as self-protection. At this point the store

personnel typically got quite irate, all of which was captured on the head cam. As it was captured, this video was wirelessly transmitted back to his computer, parsed and edited by algorithm, and automatically posted to his Web site. (2)

Can a camera help you make sequences? Barbara Barry recently received a PhD for her work on the "mindful camera." The crux of this experiment was to explore how a partnership might be created between a person and a computer-augmented camera that helps us think about how to create a sequence. Today, computers are blind, deaf and dumb; humans, on the other hand, are able to make sense or meaning out of what they see and hear. In order for the camera to help us in our decision-making, it must be able to understand the situation we are in and its narrative potential. If the computer has access to large databases of common-sense knowledge about the world, can it help us build a particular story? Can we collect common story sense? How can we help the machine reason using these resources? In the "mindful camera" system, the human videographer annotates her shots as they are taken in the field. The computer processes these annotations and returns some related information that might help the videographer think about the situation. A critical contribution of this research was to recognize that the videographer has a "cycle of reflection" which begins after they complete a shot and ends with the decision to take the next shot. This cycle begins with a period of evaluation: "Did I get the shot I wanted?" or "How does the shot I got contribute to the story?" This is followed by a period in which a machine can interact with the videographer, first by accepting an annotation and then by providing suggestions that expand upon the story potential. Both the annotations and the suggestions are valued as metadata, a subject to which we will return. (3)

Another experiment that increasingly intrigues students concerns instrumentation that captures everything that happens from multiple camera perspectives. Often this involves replacing human operators with "smart camera" technologies that allow multiple cameras to self-organize and cover sports events, musical performances, or other situations from multiple points of view. This vision was articulated many years ago by a student who thought that in the future multiple semi-autonomous cameras would be able to emulate the style of Richard Leacock, the documentary filmmaker. These cameras should be able to move about a space full of people and collaborate amongst themselves to achieve an artistic end. The student expressed the idea that this approach could free Leacock to do what he does best, which -- according to the student -- was to "engage with his subjects." To me the student's thinking was misguided not only because this might not ever be technically feasible but also because the student did not acknowledge that Leacock's passion for making movies was itself a passion for discovery and experimentation. The whole fun for Leacock has to do with examining a human situation through a camera lens and coming to a new understanding of it. The autonomous camera solution would destroy the very type of engagement that makes Leacock's work interesting: that is, the way in which an individual's cognitive and emotional engagement shapes a point of view

and drives their filmmaking. I cannot imagine that Leacock would be in the slightest bit interested in hanging out while autonomous cameras whirred around him.

At the time it was proposed, the autonomous system described above was not feasible. Last year a student, Peter Sand, came close to realizing the above scenario. He had also spent time thinking about how to make motion rigs for video. I was the ideal person with whom to do an independent project as I was concerned about what was being captured and how it might communicate a "sense of being there."

When Peter first came to me, he was interested in developing a camera rig that would lend itself to both manual and computer control. The rigging could range from a basic pan/tilt unit to a large jib or stabilization device, depending on available time and resources. This idea was: once the rig was built, he would be able to experiment with driving the camera based on real-time computer vision and/or interactive tangible interfaces. He also wanted to develop algorithms to autonomously record and edit some of his footage into a music video.

As is frequently the case, this project underwent some adjustment. When the producer of the band Peter wanted to work with saw what Peter had in mind, he decided that the band would gain more benefit from a detailed documentary record of their various activities. The revised project involved developing a ubiquitous recording scenario that would result in as complete a catalog as possible of the bands various activities: composing, practicing, touring.

First, Peter developed the individual recording units: a small computer equipped with a motion detector, a microphone, and a mini-DV video camera. The idea was that when the computer detected motion or sound, it would begin recording video to disk. Later the system would go back and recompress the footage to allow for more recording. The units were networked so that they could back up files on machines units that were not in use and send commands from one to the other to start/stop recording.

Peter began by placing four units in the practice area and two cameras near the piano in the songwriter's home. Footage streamed in at a great rate but, as expected, much of the footage in the practice space was not framed in a very compelling way. The cameras were in the corners of the room, which meant that the images generally included backs as well as faces. In addition the disc arrays could not keep up with the data flow. A nice try, but there is still lots of work to be done before autonomous recording of "a day in the life" can be realized; the "surveillance camera" approach to filming produces results that are aesthetically lackluster and intellectually unfocused. (4)

Currently, Professor Deb Roy is trying a similar experiment, this time with more gear and a different intent: he is interested in understanding how humans acquire

speech and gesture. A new baby recently came into the Roy household, which is now full of surveillance cameras and microphones. The first issue is still one of resources: huge disc arrays are needed to store the copious amounts of video captured. Once the issues associated with the capture and storage portion of the project have been realized, Deb can begin that quest for finding those relevant moments in the data. Good luck Deb! (5)

Summary: When we have decided what to capture, we need to determine what gear is required and how it should be deployed. Often, a hand-held Digital Video camera equipped with a microphone is the best fit for our purposes. Human operators quickly learn to think about sequences and, with some practice, discover how to manipulate the camera to their liking, in ways that allow them to edit later. An important refinement to this set-up will be the ability to easily annotate footage with useful information as it is shot in the field. A further and important improvement will come when we are able to create a system to think with us while we are shooting. The "Mindful Camera" begins to explore this space. (Barry.. ref).

However, there are many life situations that we might want to capture that require more than a single hand-held camera. Television has already tackled sports, musical performances and social crises using multiple cameras controlled by human beings. As we develop autonomous computer-controlled cameras, we can begin to capture more complete recordings of a range of phenomena. This leads us to suggest that before deciding what gear to use, it is important to articulate the intention behind the recording. Is the purpose of the recording to entertain, to provide evidence, or as a means of extending our memory and perceptions?

3rd Problem: mindful video: of metadata and presentation

Rich-media storytelling depends on the ability to transfer the synthetic sense memory from the still image or video representation to an audience and stimulate their imagination, understanding, and empathy. The making and experiencing of stories is a human pleasure and a privilege. Every time we engage deeply with a still picture or movie, we add to our experience of the world and our understanding of ourselves.

The complexly nuanced sensory information in rich-media representations provides a surrogate experience that resonates with our own sense-memory at different levels: a fleeting similarity between the impression of light and shade; a profound connection between what we find in the imagery and something we have experienced or other stories we have encountered. As we engage more closely with images and movies, we may pause to consider: how is it that the many people who see this movie interpret it in such similar ways? What mechanism do we use to parse the meaning of a movie? Why is it so difficult to describe our understanding in words? One thing is certain: the human being has

developed an elaborate mechanism for interpreting the flow of images, sounds and story. We parse the movie into subsets of correlated meaning that correspond with the structure imposed by the filmmaker. We know instantly when a new scene has begun and we then allow our brains to process with closure the meaning of the scene that has just past.

Video stories are constructed by selecting and ordering one or more shots and sounds into a meaningful sequence. Sometimes the editor is seeking the next shot in a sequence. Sometimes she is selecting the shot that will set up and begins a new scene. The editor is always concerned with what arrangement will create the right meaning.

In 1983, I began a research endeavor that has lasted over a quarter of a century. How can we make an editor in software? Before expanding the discussion of how, let us analyze why. There are two aspects: the first is to search for and find material that is relevant to the story; the second is to select the next shot or next sequence to be shown.

When we are collecting media, we may build our collection around something in particular or around nothing in particular. If this collection grows to a reasonable size, the problems of naming, searching for, finding and sequencing becomes difficult. While we may retain a vague memory of a shot or of an event we attended with our camera, our exact memory of all of its specific visual and sound details quickly grows dim. In order to find the shot we use “memory handles:” for instance, it was during the summer when the family was all together, it was someone’s wedding, Aunt Flora was there, etc.

Obviously this task is ever more difficult for someone who is unfamiliar with the collection and lacks any emotional ties or “inside information” about the footage even though she may view it and find it meaningful.

In building the editor in software, metadata becomes our accomplice.

How can a picture be described?

The old adage “A picture is worth a thousand words” proves its truth as we search for effective ways to represent image and sound media. These “thousand words” are often stripped down to the most useful set of Metadata descriptions.

Wikipedia begins its description of Metadata with the derivation “(Greek: meta-+ Latin: data "information"), literally “data about data, information that describes another set of data.” (6) It goes on to offer the common example of the library catalog card that includes ID information for the book, its author and publisher, a general cateqorozation of its subject matter, its size, and a pointer to its physical location. Another equally useful example is a computer file system which automatically generates and updates metadata associated with a file: its size, the

date and time it was created, when it was last modified, its actual location in computer memory and the “filename” provided by a human.

Metadata comes in many forms. Some can be generated by the camera itself, such as: time code, f-stop, lens focus, GPS location coordinates, compass orientation and so forth. Historically, the most valuable metadata is added by humans after the time of capture, such as keyword descriptions and text paraphrase that highlights the important “who, what, where, when, and why” of the shot’s content. Some metadata can be generated by signal-analyzing machines: extracted patterns and “blob tracking” are examples of this. Other metadata accumulates as the material is used: its position in an edited sequence, how many people have viewed it and when, what is most often viewed after it, etc.

Metadata can be used by human editors and computational engines (with or without the intervention of humans) to locate specific shots, to suggest similarities in collected materials, or to make a best guess at “what shot to show next.” However, none of these methods have to date provided a reliable, scaleable approach to the “next shot” problem for large collections of synthetic memory recordings. In the paragraphs that follow we survey a variety of metadata experiments.

The value of metadata largely depends on whether and how efficiently it can be generated and used by a computer or by a human. For instance, in traditional film editing -- when the picture is on reels of movie film and the sound is recorded on separate reels of sprocketed magnetic stock -- the processing lab periodically stamps human-readable “edge numbers” on both. By matching up the numbers, sound is kept in sync with the picture. However, such a system would be useless for video; so, in the 1960’s SMPTE time code was invented and standardized. This is a form of metadata that is recorded into every frame of video footage shot by professional cameras; part of it assigns a unique number to each frame, and part of it is available for other useful tags. Most importantly, it is machine-readable, allowing automatons to perform frame accurate video edits on a repeatable basis. These numbers also enabled humans to reference and locate content they have described in a content or continuity log.

For some applications, knowing where an image was captured can be an extremely useful aid for finding, selecting, ordering and manipulating it and related images. In 2004, researchers in Marc Davis’s Garage Cinema group at the University of California, Berkeley explored the idea that someone visiting a place might want better pictures of the site than she was able to capture with her cell phone’s built-in camera. The researchers built a system that exploited the capabilities of GSM/GPRS-enabled cell phones, which can stamp the images they capture with GPS (geographical positioning system) location coordinates and date/time information. Using image-analysis and inference techniques, the system makes a “best guess” about what the user was trying to photograph,

compares the captured image with other pictures from the same location, and offers other examples from a database of collected images. If the user likes a suggested image better then the one she just took, she can add it to her collection. (7)

Another approach to metadata that has proven relatively successful is the addition of keywords. Keywords describing selected aspects of the video and audio are added explicitly to media in order for a system to search and find data useful shots and sounds more effectively. Certain computational techniques can exploit keyword annotations to automatically select an "appropriate" next shot.

Over the past 15 years, two orthogonal approaches to keywords have been developed for audiovisual collections. In the early 1990's, the Getty Museum in Los Angeles spearheaded an effort to create an "expert" taxonomy of keywords for museum collections. The idea was to standardize the vocabulary used to tag art objects so that meaningful, accurate searches could be made across collections by posing a well-formulated query. In other research domains, proponents seek more fluid mechanisms unconstrained by a rigid, commonly agreed-to vocabulary, ones that allow the human to use free text and the machine to reason about the text in a storied way.

What do we mean by "a storied way?" Must "storied" be different for different authors, different content sets, different audiences? Or are there some general principles that we can articulate that maximize the "storiedness" for a computational collection-based media environment?

Among other things, to be storied the "next shot" must act in concert with previous shots to extend the recipient's sense of meaning. Continuity becomes one metric of storiedness – that is, the "next shot" should not accidentally break the recipient's built-up expectations of who is in the world or how the world works.

In order to arrive at an appropriate level of storiedness, the computational mechanism needs to act on something that can produce continuity. In the "JBW: A Random Walk through the 20th Century" project -- an on-line interactive biography -- we built a collection of writings of Jerome Wiesner, former President of MIT, and video segments in which people told stories about their interactions with this man. The interface allows the audience to affect a state machine by selecting various elements in the concept map interface that represent keywords that have been arranged into classes of who, what, where when. By manually selecting a keywords, the audience alters the weighting of the keywords and therefore affects the play-out of segments in the collection. This sets up a dynamic that allows certain axes of meaning to linger while others shift which insures a plausible appearance of storiedness. (8)

With a small number of media elements and a semi-coherent collection, it is fairly easy for an author to standardize keywords. However we need to consdier the case of a very large evolving collection to which many authors contribute. In this environment, both the author(s) and viewers will be more comfortable using free text to describe their interpretation of a video segment. In the late 1980's, Ricki Goldman developed a large database of video focusing on children and learning. In this case, video was used to explore children's thinking. It was part of "Project Headlight", a 5 year undertaking by Seymour Papert to explore constructionist learning at the Hennigan School in Boston. Ricki built a system that allowed researchers to access the video material and to add reflective commentary to any part of the video. In this implementation, no advanced computer methods were used to parse and evaluate the commentary, rather the activity was person to person. (9) A similar limitation exists in blogs today. In current research at the Media Laboratory, we are bringing the power of some analytic tools, including we "WordNet" (10) and "ConceptNet" (11) to bear on this problem in hopes of creating a partnership with the computer that would allow us to find appropriate nex shots in very large video collections.

While keywords and descriptive commentary are usually added to media after shooting (and are generally the product of human judgement and labor), many researchers feel that machines should be able to extract meaning from the video stream itself. Over the years there has been some interesting research in this area such as "p- " which tries to find similaries between faces in live situations or pre-recorded video; this type of analysis fails when large variations of light, scale and framing occur. For this reason the results have not been sufficiently robust for our purposes. (12) On a different tact, one promising experiment some years back involved looking for laughter in the sound track, in the belief that laughter would signify interesting material. (13) While fun and promising, this method is also somewhat narrow and needs to be combined with other techniques if it is to be of significant use for everyday video collection.

A final type of data we might use to enhance the "storyness" of the "next shot" is user-provided data (as opposed to the usual author-provided). The web site "Flickr" currently allows users to add images, annotate their own and other peoples images, and leave user "tags" that reflect how an individual might think of using the image. Flickr uses these tags to rank images and assemble albums. Based on millions of users interacting in different yet similar ways, headings such as "my images", "most popular" and thematic albums begin to map this image space in new and lively ways. Still, these mechanisms do not yet move the content into a more "storied" framework; rather, the human interactions around the site seem to carry the storiedness. (14)

Conclusion: the question of the channel

The human mind communicates with the world around us through biological sense-channels. Technology allows us to capture synthetic sense-memories,

mould them into intentioned objects or streams that can deepen, widen and make more communal our biological memory.

What I hope I have now made evident is that our desire to engage – to make, manipulate and consume these memories – must be met with a complex commitment of time and mental activity. It is therefore not unusual for us to search to articulate how such a repository of captured imagery will be useful. Useful to whom? To ourselves? To our children and grandchildren? To our organization? To a world-wide audience?

Usefulness generally requires that the memory be shared – with our selves and others, now and in the future. This brings us to the idea of the channel.

In person engagement occurs face-to-face using our biological sense channels: our perception of the the signals emitted by others are subject to laws of physics, to the state of our senses, to our continuous conscious presence in the world, and to cognitive abilities to parse, reflect and interpret. Augmented by hardware and software , our synthetic sense memory can reach us in different ways: "push" technologies such as broadcast television and some advertising on the WWW are generally agnostic to the party on the other end of the pipe, where "pull" technologies (most WWW applications) deliver material based on request. Preparers of content make content for broadcast (the same content goes to millions of viewers) or for narrow cast (my content may only interest a few people; how can I find them?)

More than ever before, the narrow cast option has infinite potential. Any one can put a dense individually-authored collection of media up on a channel, a blog for instance. With the help of metadata and a small number of filter types, this collection can be infinitely parsable, so that no two people see the same selection of clips in the same order. Take for instance, my home movie sequence collection: any member of my family can search through the collection and see those sequences in which they are present. If Caroline is the viewer, for instance, she might choose to remain true to the "Caroline channel". With the addition of our spreading activation approach used in JBW (8), any visitor can pick a starting point and navigate through the collection. Rarely will any two people experience the same story, although everyone will probably get a similar sense of family. Each individual viewing path can be considered a channel.

To go back to our earlier discussion, this phenomenon of having almost an infinite number of channels suggest that, in today's world, the combination of what you shoot and what metadata exists for that footage defines what channels become available. This is a critical realization if we are going to take on a large recording effort such as documenting a school.

While rules of the game may need more definition, let us assume for the purposes of discussion that that anyone can record new material that concerns

the school. Let us also assume that there is a rule that anyone who captures such material is bound to archive it in the school's archive. By archiving we include the requirement of some metadata annotation that identifies the name of the person recording, the date, the circumstances and other annotations that are deemed appropriate. This idea raises the following questions that we can think about together in class:

* Why should we collect video? What channels are we collecting video for?

This leads us to consider additional questions:

* Should there be a minimum amount of video that is pre-defined and regularly captured?

* Do any circumstances require the use of multiple cameras?

* Should the archive handle rushes or edited segments of video?

* How can the school afford and sustain all resources: gear and human time?

* What kinds of metadata should be required for each video segment? What technologies do we need to make use of that metadata?

* What are the rules governing distribution and use of this video?

Footnotes and references:

1) Richard Leacock has discussed this problem of how to shoot performance extensively. He touches on the problem briefly in A Search for the Feeling of Being There (1997) which can be found in the essay section of his web site. The first person to really address the problem of shooting an orchestra for live network distribution was Bill Cosell at WGBH http://www.richardleacock.com/leackessays.html#A%20Search%20for%20the%20Feeling%20of%20Being%20There

2) Steve Mann, More on art (invited plenary lecture at Ars Electronica, along with a week long performance piece there called "Sicherheitsglaeser"): http://n1nlf-1.eecg.toronto.edu/sicherheitsglaeser/

3) Barry, B. (2005). Mindful Documentary. Massachusetts Institute of Technology, Ph.D. thesis.

4) Peter Sand, Technical Report, May 2005.

5) Deb Roy

http://www.media.mit.edu/cogmac/projects.html

6) http://en.wikipedia.org/wiki/Meta_data

7) Marc Davis and Risto Sarvas. “Mobile Media Metadata for Mobile Imaging.” In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2004) Special Session on Mobile Imaging in Taipei, Taiwan*, IEEE Computer Society Press, 2004.

8) http://ic.media.mit.edu/projects/JBW/

9) ricki goldman-segal, …

10) wordnet described: http://wordnet.princeton.edu/w3wn.html

11) concept net described: http://web.media.mit.edu/~hugo/conceptnet/

12) Sandy Pentland, face recognition , see video entry under “faces” on http://web.media.mit.edu/~sandy/

13) Lockerd, F. Mueller, “LAFCam: Leveraging Affective Feedback Camcorder”, ACM Conference on Computer Human Interaction. April 2002.

14) flickr in action; become a member see how it works, and evaluate relevance http://www.flickr.com/