Computer-Supported Collaborative Video Analysis

Roy Pea, Robb Lindgren and Joseph Rosen
Stanford University, Stanford Center for Innovations in Learning

Video can serve as a powerful medium for analyzing interactions involved in learning activities, for capturing records of teaching for uses in professional development, and for learners to construct or interact with videos expressively, but there have been many barriers to its collaborative uses. The DIVER Project is tackling core problems in advancing computer-supported collaborative video analysis. DIVER establishes a unique video platform for users to control a "virtual camera window" on a computer for *guiding noticing* to parts of a video of interest as the video is playing—then to text annotate these moments, and publish them on the web for further collaborative analysis. Every DIVER user—researcher, teacher, or learner—can thus express their visual and interpretive point of view on a video, providing an important and accessible tool for *making a difference* in our research and education. Ongoing studies of learning research and higher education applications using DIVER are discussed.

Introduction

Video has been an unnatural medium to support collaborative activity in a way that provides persistent records of the insights that collaborators develop together. Text provides for collaborative writing, but when researchers or other learner groups want to work with video as a medium for interpretation, sharing, commenting, their collective work is not well enabled. Our challenge is making computer-supported collaborative video analysis (CSCVA) into a commonplace collective practice, whether face to face or in distributed groups connected with mediating technologies. We are a long way from collective practices of video analyses in the learning sciences or for education. We describe the socio-technical design problems facing CSCVA, in order to explain our rationale for and experiences with DIVER as a platform to support video collaborations for several ongoing scientific and educational activities. We then consider how workflow scenarios with DIVER are meeting the CSCVA socio-technical design challenges with which we began.

Background and Significance of the CSCVA Problem

The advancement of augmentation tools for human activities has built on several decades of socio-technical design theory and practices. Scacchi (2004) notes how: "Socio-technical design is concerned with advocacy of the direct participation of end-users in the information system design process. The systems include the network of users, developers, information technology at hand, and the environments in which the systems will be used and supported. The process includes the design of the human-computer interface and patterns of human-computer interaction." With this view in mind, we ask, what are the primary socio-technical design issues for CSCVA that heed these concerns? We characterize seven primary socio-technical issues that constitute core challenges affiliated with CSCVA. Virtually all the concerns to be outlined relate to the fundamental problem of *coordination* of attention, interpretation, and action between multiple persons. Clark (1996) described as "common ground" what it is people seek to achieve in the work they do to coordinate what they are attending to and/or referring to, so that when comments are made, what these comments refer to can be appropriately inferred, or elaborated. In the learning sciences, "common ground" is usually used to examine collaborative or teaching learning discourse and pointing, bodily orientation and joint visual regard to the focus of a conversation that is being analyzed for studies of learning or teaching (e.g., Barron, 2003). But it is not sufficient only to focus on the non-technology mediated aspects of common ground—once we look at inscriptional systems (e.g., Latour, 1986) that lay down and layer symbolic records such as text, diagrams or other paper-based representations, they, too, become a focus of pointing and joint visual regard, and introduce new problems

as transient referents. One may want to refer to an earlier moment when only a part of a mathematical diagram was present, for that is what one wishes to establish common ground around. This common ground concept must extend from the face-to-face situation of co-presence and its static properties to the situation of distributed conversationalists and the dynamics of representations, as when these representations are computer-enabled (e.g., Pea, 1994). One often needs to refer to specific states of information display when using computer tools, so establishing a common ground for discourse and sense-making in a digital realm is important, too, including the dynamic medium of video. Considering these concerns, our seven socio-technical design issues for CSCVA are:

1) The problem of reference. How is it that I as someone analyzing video can refer to a specific time-space region of a dynamic video record in a way that 'lasts' beyond the here and now? In the here and now, I can point my finger to a display on which a video record is playing, encircle the topic for my comment, and thereby highlight what I will be commenting on. Traditionally, the video medium has not provided the capability for a watcher of the video to point in such a way that *records* the locus of reference for a time-shifted or space-shifted (remote) audience of the watcher's referring act. It is noteworthy how this is important even for an individual analyzing a video as well; absent a record of what I have pointed at, I may later not be able to recall what aspects of the video I found interesting, problematic, compelling, or whatever, or to which a comment I wrote down can be attached. So solving the video reference problem can also serve as a personal memory aide, not only a collaborative one.

2) The problem of attentional alignment. The related problem of co-reference, or 'attentional alignment', has to do with how it is that I as someone analyzing video and engaging in a referential act to some part of it can establish that the person(s) to whom I am addressing this referring act is attending to the video segment that I refer to. Goodwin (1986) highlights gesture's deictic uses in efforts to organize mutual orientation. When the conditions for achieving such co-reference have been secured, I have been successful in establishing attentional alignment. The reason that this coordination is important is that any consequent dialog about my video referent can lead to misunderstanding and other conversational troubles if my listener believes I am referring to something other than what I intended to refer to.

3) The problem of creating video "immutable mobiles." In characterizing the power of written texts, Latour (1986) developed the influential concept of inscriptions as external representations of ideas that serve as "immutable mobiles" with these key properties: 1) Inscriptions are mobile; 2) They are immutable when they move; 3) They are made flat; 4) The scale of the inscriptions may be flexibly modified; 5) They can be cheaply reproduced and spread; 6) They can be reshuffled and recombined; 7) One may superimpose different images of totally different origins and scales; 8) They can be made part of a written text; and 9) Their 2-D character allows them to merge with geometry. Although his concerns were to provide a novel explanatory account of how science and technology took hold so powerfully, Latour's theory has had considerable influence in the digital documents world (e.g., Levy, 2001), and its implications for conceptualizing video as an inscriptional medium are important. How shall videorecordings become immutable mobiles? (Stevens & Hall, 1997).

4) The problem of effective search, retrieval and experiencing of collaborative video work. The work life of video analysis with videotapes was challenging enough, with the physical tape media to index and store, possibly in many different versions as derivative tape collections were made of interesting moments. If we can solve the problems of pointing to video, establishing attentional alignment, and creating video as an immutable mobile medium, we generate a few new problems. As users become technically enabled to point to, annotate, and foster attentional alignment to portions of a videorecord, they will generate a vast array of persistent video-plus-pointing-plus-commentary digital objects. How will users be able to effectively solve the information retrieval problem, and quickly find what they want and experience these video-anchored interpretive acts, the moments that matter to them?

5) The problem of permissions. Access and control issues turn on appropriate assignment of permissions, with two broad classes of situation at hand. One concerns video analyses when we work

with video that may have sensitive human subjects conditions. Research participants (or the parents of minors) in from institutions funded by federal research grants provide informed consent to specific conditions of use for research videorecords of their activities. These consent forms and procedures for ensuring confidentiality of subject data, or other terms of consent (such as in what contexts the video can be shown) are reviewed and approved by Human Subjects Institutional Review Boards in a researcher's institution. For this reason, it is important to have only approved individuals view such video. In a second situation, there are issues concerning digital rights management, common for film or television works but also applicable to user-generated video content. One may wish to allow only certain individuals to view video, to create video annotations, or to make remixes of video assets.

6) *Establishing a productive workflow of collaborative activity.* This issue is really about how to best tap the collective intelligence of a group engaged together in collaboratively analyzing video recordings. In canonical interaction analysis methodology for group work on video (e.g., Jordan & Henderson, 1995), there has tended to be a hierarchical rather than heterarchical organization of analysis: a group leader controls the video selection and play, picks participants in the physical meeting to make observations, audiorecords the group's work, mines the groupwork audiotape, and then creates or directs the creation of a video analysis informed by the deliberations of the group. In such a workflow there are many idle times when observations or contributions from group members are not being tapped or recorded for comparison and reflection, and multiply rich interpretive accounts that could be developed from the collective set of insights of the group are not developed. What new kinds of research activity structures are opened up with CSCVA which are distinctive from face-to-face video analysis?

7) *The problem of establishing coherent multi-party video-anchored discourse.* Consider a face-to-face conversational interaction, as in a seminar, when the rules of discourse that sustain turn-taking and sense-making as people converse are familiar. Discourse analysis and studies of conversation have made great progress in uncovering the systematics of turn-taking, speech acts, and accounting for the semantic and pragmatic coherency of discourse across turns and speakers (e.g., Heritage & Goodwin, 1990; Heath & Luff, 1993). Intrinsic to these social activities in work meetings and academic seminars are uses of media that include paper used by individuals for note-taking, a whiteboard for writing and drawing, and increasingly, a public screen for displaying computer presentations. In an academic setting, video, film and audio recordings may also be played. Traditionally these are used asymmetrically, as the facilitator/instructor prepares these records to make a point or to serve as an anchor for discussion, and controls their play during the discourse. Computer-facilitated meetings for doing video analysis—where each participant has a computer and is network-connected with other participants and to external networks for information access, search and retrieval—bring new challenges beyond non-technically mediated meetings in terms of managing a coherent group discourse.

Illustration of Troubles to be Resolved

This paper attempts to make these issues concrete by discussing them in terms of ongoing efforts to engage in two novel practices with face-to-face and distributed CSCVA: (1) in several undergraduate *courses*; and 2) for learning sciences *research* concerning informal learning.

Undergraduate Courses Utilizing Video

Over the past year we have worked with a number of college-level instructors at large researcher universities who had all in the past employed video-based discussions into their lessons. These courses included a film studies course and a Japanese language course at a West Coast university, and a film production course at a midwestern university. In each setting the goal was to present students with one or more video artifacts and have the students generate insights using an analytic frame provided by the instructor. For example, students in the film studies course looked at two video clips, the "Crispin's Day" speech in the film adaptation of Shakespeare's Henry V from the 1989 version directed and played by Kenneth Branagh, and from the 1944 film version of the same play directed and played by Sir Laurence

Olivier. The instructor's objective was to have students comparatively analyze how the same text was translated to film by two different directors/actors. Students previously had to make the comparison for this assignment by reconstructing the film events from memory for a written essay. In this scenario, students are faced with the reference problem—in making their argument they have no way to explicitly refer to an important aspect of the video film clip (e.g., an aggressive gesture by Branagh at a point where Oliver is subdued) that they may have noticed upon first viewing.

In the film production course students are also asked to analyze video clips, but in this case the clips are those created by fellow students. Individuals are responsible for making their own experimental film available for others to view, and for providing feedback on at least two other student films. How valuable one student's feedback is for another depends largely on the issue of attentional alignment—how effective is the critiquing student in conveying to the film's creator the focus of the critique? For example, one student shot a music video for his film project and another student noted: "This is my favorite sequence. The transition between band members works well with the music and beat." Absent a record of what precisely the critiquing student was looking at when she made this comment, it carries little meaning. The film production course also exemplifies the problem of search and retrieval. As the critique of other students' films is a class assignment, the course instructor must be able to efficiently access the corpus of comments, sorted by author or target film, so as to judge the quality of the feedback.

In the Japanese language course the instructor's desire is to engage her 10-12 students in a discussion about the dialog styles used by the actors in a set of video clips she has collected over the years. Specifically she seeks to elicit insights from her students that will illustrate a sophisticated understanding of conversational Japanese; thus she uses a diverse set of sources from pop culture artifacts (e.g., Japanese soap operas and anime) to videotaped interviews with native Japanese speakers. Although these insights are not formally assessed, it is important to the instructor that all students contribute to the discussion and that these insights build off one another. Structuring a productive multi-party discourse of this type centered around a number of video records can be difficult for an instructor to accomplish given the lack of precedent for video anchored conversation and the requisite patterns of discourse.

Learning Sciences Research: Collaborative video analyses

In our NSF Science of Learning Center (LIFE), we are studying informal learning of mathematics in family situations. Our research group of two faculty and three graduate students is investigating the contexts and activities in which middle school age learners and their families engage in mathematical problem solving. In interview sessions and observations lastly roughly two hours, we are working with over 30 families that represent California's diversity to identify the contexts and situations that families participate in which serve as locations for mathematical learning and practice. We digitize our video records and seek to develop coding categories and analyses that foreground the nuances of family math in situ. We seek to describe the resources family members use for recognizing and solving problems, characterize the structure of their mathematical activities, and analyze the social conditions and arrangements for their family-based mathematics practices. We used a semi-structured interview protocol organized around mathematically relevant contexts to center on activities that most families engage in, while allowing families to give us their particularized versions of how they accomplish each life task (including technology use, and systems of representing mathematical relationships). For our analytic work, we wanted to do both face-to-face meetings reviewing videorecords, and independent work that can contribute to the group collaboration whenever we can access the data and build interpretations and coding activities concerning it. We also will be doing collaborative analyses of our data with researchers at U.Washington, UCSC, and other institutions. We need good solutions to the seven socio-technical design challenges to CSCVA to make this work as productive as possible.

THE DIVER Software Environment for Video Collaboration. DIVER is a software environment first developed for research uses of panoramic video records (Pea et al., 2004). As we have developed a web-enabled DIVER allowing for distributed access and annotation of digital video records from consumer digicams, our focus has shifted to supporting collaborative video analysis and emerging prospects for "digital video collaboratories" (Pea, in press). We are putting DIVER to work and evolving its capabilities in support of collaborative video analysis for many research and educational activities. We call the central work product in DIVER a "dive" (as in 'dive into the video'). A dive consists of a set of XML metadata pointers to segments of digital video stored in a database and their affiliated text annotations. In authoring dives on streaming videos via any web browser, a user is directing the attention of others who view the dive to see what the author sees; it is a process we call *guided noticing* (Pea et al., 2004). To make a dive using DIVER a user logs in and chooses any video record that has been made available in the searchable database. The video selected can be viewed using standard video controls. As the video plays, a user can manipulate a virtual camera viewfinder (the yellow rectangle in the figure) on top of the video to focus in on a specific area of interest. By clicking the MARK button, the user saves a reference to a specific point in space/time within the video and places it within a data container—a single panel that resides inside the DIVER "worksheet" in the webpage and signified with an image thumbnail. Once the mark has been added to the worksheet, the user can comment on that mark by entering text in the panel. Panels can also be created by clicking on the RECORD button, creating a pointer to an entire segment of the video and storing the path taken by the virtual viewfinder during that segment. Like a MARK, a recorded clip can be annotated by adding text inside the respective panel on the worksheet. The DIVER user can replay the recorded video segment or see the recorded mark by clicking on its thumbnail image.

Assuming they have appropriate permissions, multiple users can access a dive simultaneously, with each user able to add new panels or make comments on a panel that another user created. Users are notified in real-time when another user has made a contribution to the dive and they can view any changes by clicking on the update button. Thus, users may be either face to face in a meeting room, or connected to the same webpage remotely via networking, as they build a collaborative video analysis. In principle and in practice, there is no need for the users to be watching the same portions of the video at the same time; as the video is streamed to them through their web-browser, they may mark and record and comment at their own pace and as a function of their own interests. Collaborative video analysis activity using DIVER can be as planful or emergent as participants choose to make it; constraints on focus, intent, duration of sessions, and so on are not built into the technology but a matter of negotiated social practice.

CSCVA Workflow with DIVER: Addressing the socio-technical design issues

With a brief sketch of DIVER, and these examples where research or learning groups have been using DIVER in support of collaborative video analysis, we now may reflect on the workflow steps and the activity systems in which the DIVER technologies are playing instrumental roles. We have been learning a great deal about the unique affordances of this video analysis platform for the nature of such work, and unearthing new challenges.

1) The problem of reference. DIVER provides a virtual camera viewfinder for a user to inscribe a video region of interest as their referential act (Stevens et al., 2002 use a pointing gesture). Text annotation is used for free-field text interpretations or coding categories to the referred-to video content. Virtual pointing to a video region and affiliated annotation provide the communication infrastructure for a dive author to make *link-addressable* references to the dynamic medium of video in their conversations.

2) The problem of attentional alignment. The method of virtual pointing to a circumscribed part of the space-time continuum of a video record enables distributed users to focus their attention on the same regions of the video for their interpretive work. With DIVER's web-based methods, users can align attention to the parts of the video that matter to them for their discourse whether they are synchronously or asynchronously connected, as the points into video streams are persistent XML metadata.

3) The problem of creating "immutable mobiles" from video recodings. Latour's concept of written texts as immutable mobiles reviewed earlier has considerable applicability in considering how a dive establishes a video immutable mobile. The digital inscriptions provided by the lightweight metadata of a dive are mobile, immutable when they move (in that they preserve their character across locations, platforms, browsers), are made "flat" (2-D), can be flexibly modified in scale (through projection), can be cheaply reproduced and spread (thanks to Internet standards), allow one to "superimpose different images of totally different origins and scales" (in their digital document forms), can be made part of a written text (through hyperlinking to dives), and may be merged with geometry thanks to their two-dimensionality.

4) The problem of effective search, retrieval and experiencing of collaborative video work. The time dimension of video needs to be unlocked. With any volume of video, this is an enormous problem and a barrier to greater video use. DIVER provides Google-like search for any term or phrase used in the full text of annotations, title, or user name. A DIVER search returns a list of dive panels with the search term or phrase highlighted and the video keyframe for that panel. In terms of retrieval, clicking on such a list item or keyframe opens up the affiliated dive and enables the user to view the precise video regions in space/time to which the searched-for terms were applied. The value of the metadata tagging of a user community for research videos will grow tremendously as multiple researchers work with a data set and develop cumulative analyses across projects. In the informal math learning work, we are finding it easy to compile the accumulated annotations of our multi-party research group on specific phenomena such as uses of props, symbol manipulation, emotion terms, and the like and then to create composite codebooks that are dives resulting from all exemplars of a given type, identified across multiple researchers, and now able to be experienced in a sequential remix of the clips of a given type from a new Dive.

5) The problem of permissions. DIVER provides access to video records and their dives only for users granted permissions to certain rights. Administrative tools enable the formation of groups, and the establishment of whether specific video assets and dives can be viewed, copied, edited or deleted.

6) Establishing a productive workflow of collaborative activity in video analyses. Some of the most dramatic changes in DIVER-enabled collaborative video analysis concern this design problem. Whereas sequential turn taking is required in collaborative face-to-face video analysis work practices, parallel analyses can be carried out with DIVER, to cumulative effects. Multiple individuals can be streaming the same video file, looking at different parts of it at the same time, and making their dive recordings and annotations without control from the single group leader customary in video interaction analysis sessions. Stopping and starting the video being played, setting in and out points to segments to be annotated, selecting different segments to compare in an analysis, can all be carried out in parallel when DIVER is used by a distributed group of analysts. DIVER CSCVA thus shifts control from the group facilitator to individual participants, who can play the entire source or can use the cues from one another's ongoing analyses to focus their attention on subparts of the video. In relation to scientific inquiry processes, there is the added benefit that, as one posts conjectures as DIVER annotations concerning the interpretations of an event, multiple analysts can seek out confirmation or disconfirming evidence elsewhere in the video data records relating to that conjecture, and then post such links as either

comments in the dive of the individual making the conjecture or as new panels in that dive, pointing to other video evidence that is available.

7) The problem of establishing coherent multi-party video-anchored discourse. Our experiences with DIVER for video analysis in collaborative groups have not yet made much headway on this dimension. We have largely used DIVER in an asynchronous manner for collaborative group work of our informal family math learning video records, or face to face meetings where a projected screen is used to coordinate group attention and workgroup participants take turns by ordinary social conventions in selecting dives they would like to share or ask questions about from other members of the workgroup.

Conclusion

The DIVER system distinctively enables what we call "point of view" authoring of tours of existing video materials in a way that supports sharing, collaboration, and knowledge building around a common ground of reference. We are in the early days of documenting its uses as a digital video collaboratory platform and addressing the seven socio-technical design challenges for computer-supported collaborative video analysis.

References

Barron, B. (2003). When smart groups fail. *Journal of the Learning Sciences, 12*(3), 307-359.

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Goodwin, C. (1986). Gestures as a resource for the organization of mutual orientation. *Semiotica, 62*, 29-49.

Goodwin, C., & Heritage, J. (1990). Conversation analysis. *Annual Review of Anthropology, 19*, 283-307.

Heath, C., & Luff, P. (1993). Disembodied conduct: Interactional asymmetries in video-mediated communication. In G. Button (Ed.), *Technology in working order: Studies of work, interaction, and technology* (pp. 35–54). London: Routledge.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences, 4*(1), 39-103.

Latour, B. (1986). Visualization and cognition: Thinking with eyes and hands. In H. Kuklick, & E. Long (Eds.), *Knowledge and society* (pp. 1-40). Greenwich, CT: JAI Press.

Levy, D. M. (2001). *Scrolling forward: Making sense of documents in the digital age*. New York: Arcade Publishing.

Pea, R. D. (in press). Video-as-data and digital video manipulation techniques for transforming learning sciences research, education and other cultural practices. In J. Weiss, J. Nolan & P. Trifonas (Eds.), *International handbook of virtual learning environments*. Dordrecht: Kluwer Academic Publishing.

Pea, R., Mills, M., Rosen, J., Dauber, K., Effelsberg, W., & Hoffert, E. (2004). The DIVER™ project: Interactive digital video repurposing. *IEEE Multimedia, 11*(1), 54-61.

Pea, R. D. (1994). Seeing what we build together: Distributed multimedia learning environments for transformative communications. *Journal of the Learning Sciences, 3*(3), 285-299.

Scacchi, W. (2004). Socio-technical design. In W.S. Bainbridge (Ed.), *The encyclopedia of human-computer interaction*. Berkshire Publishing Group.

Stevens, R., Cherry, G., & Fournier, J. (2002). Video Traces: Rich media annotations for teaching and learning. *Proc. CSCL 2002*. Boulder, CO,

Stevens, R., & Hall, R. (1997). Seeing tornado: How video traces mediate visitor understandings of (natural?) phenomena in a science museum. *Science Education Special Issue: Informal Science Education, 81*(6), 735-747.