DHB: A Multiscale Framework for Analyzing Activity Dynamics

James Hollan, Edwin Hutchins, and Javier Movellan
University of California, San Diego

What conditions can facilitate rapid advances and breakthroughs in behavioral science to rival those seen in the biological and physical sciences in the past century? The emergence of cognitive science and the converging view across multiple disciplines that human behavior is a complex dynamic interaction among biological, cognitive, linguistic, social and cultural processes are important first steps. While empirical and theoretical work is rapidly advancing at the biological end of this continuum, understanding such a complex system also necessitates data that capture the richness of real-world human activity and analytic frameworks that can exploit that richness.

In the history of science, changes in technologies for capturing data, as well as those for creating and manipulating representations, have often led to significant advances. The human genome project, for example, would have been impossibly complex without automatic DNA sequencing [31, 43]. Recent advances in digital technology present unprecedented opportunities for the capture, storage, analysis, and sharing of human activity data.

Researchers from many disciplines are taking advantage of increasingly inexpensive digital video and storage facilities to assemble extensive data collections of human activity captured in real-world settings. The ability to record and share such data has created a critical moment in the practice and scope of behavioral research. The main obstacles to fully capitalizing on this opportunity are the huge time investment required for analysis using current methods and understanding how to coordinate analyses focused at different scales so as to profit fully from the theoretical perspectives of multiple disciplines.

We propose to integrate video and multiscale visualization facilities with computer vision techniques to create a flexible open framework to radically advance analysis of time-based records of human activity. We will combine automatic annotation with multiscale visual representations to allow events from multiple data streams to be juxtaposed on the same timeline so that co-occurrence, precedence, and other previously invisible patterns can be observed as analysts explore data relationships at multiple temporal and spatial scales. Dynamic lenses and annotation tools will provide interactive visualizations and flexible organizations of data.

Our goals are to (1) accelerate analysis by employing vision-based pattern recognition capabilities to pre-segment and tag data records, (2) increase analysis power by visualizing multimodal activity and macro-micro relationships, and coordinating analysis and annotation across multiple scales, and (3) facilitate shared use of our developing framework with collaborators, the wider NSF SIDGrid community, and internationally via our participation in the Rufae augmented environments network.

The work we propose builds on our long term commitment to understanding cognition "in the wild" [30], developing multiscale visualizations [5], and recent experience automatically annotating video of freeway driving. We propose to extend the theory and methods developed in our earlier work and integrate them with new web-based analysis tools to enable more effective analysis of human activity. As initial test domains we will focus on understanding activity in high-fidelity flight simulators and the activity histories of workstation usage and the process of writing. We will also evaluate a novel technique to assist in reinstating the context of earlier activities.

Our long range objective is to better understand the dynamics of human activity as a scientific foundation for design. The multidisciplinary research team we have assembled, spanning cognitive and computer science and involving existing research collaborations with investigators in learning sciences and engineering, is uniquely well qualified to conduct the proposed work, and has carefully defined a staged research approach and milestone-based management plan.

The *intellectual merit of our project* will derive from developing a multiscale analysis framework for representing and analyzing the dynamics of human behavior, integrating it with existing software tools for data collection, visualization, and analysis, and evaluating the augmented framework in selected real-world domains.

The broader impact of the proposed activity is to provide critical analytic capabilities to support research networks in areas beyond our topical research areas. In general, any research using video or other time-based records in order to document or better understand human activity is a potential beneficiary of the proposed work.

Our work will be disseminated through publications and a community-oriented website in accordance with University policy. Regular video-conferencing interactions will serve to communicate with current collaborators, extend interactions to a wider community, and encourage continued growth of a research community with shared interests in understanding the dynamics of human activity.

A.1 Introduction

What conditions can facilitate rapid advances and breakthroughs in behavioral science to rival those seen in the biological and physical sciences in the past century? The emergence of cognitive science and the converging view across multiple disciplines that human behavior is best seen as a complex dynamic interaction among biological, computational, cognitive, linguistic, social and cultural processes are important first steps. While empirical and theoretical work is rapidly advancing at the biological end of this continuum, understanding such a complex system also necessitates data that capture the richness of real world activity.

A new generation of inexpensive digital recording devices and storage facilities is revolutionizing data collection in behavioral science, extending it into situations that have not typically been accessible and enabling examination of the fine detail of action captured in meaningful settings. This is important because to understand the dynamics of human and social activity, we must first understand the full context of those activities and this can only be accomplished by recording and analyzing data of real-world behavior. While such data are certainly needed, more data cannot be the whole answer, since many researchers already feel that they are drowning in data. Data without appropriate theoretical and analytical frameworks do not lead to scientific advances.

Fortunately the revolution in digital technology can be coupled with exciting recent developments in cognitive theory. While these developments also heighten the importance of understanding the nature of real-world activities, they are in addition beginning to provide an analytic framework for understanding how cognition is embedded in concrete contexts of human activity. Cognition is increasingly viewed as a process that extends beyond the skin and skull of the individual [16],[17],[31], [28],[45],[47], [49],[51]. This shift in framing the unit of analysis for cognition introduces a host of previously overlooked cognitive phenomena to be documented, studied and understood.

Rich new digital data sources coupled with this shift in theory promise to advance understanding the links between what is in the mind, and what the mind is in. For example, just as widespread availability of audio tape recording supported the development of conversational analysis [22, 29, 50] and the ethnography of speaking [3, 24], the advent of inexpensive digital video is starting to have a fundamental impact on cognitive science. The ability to record, view, and re-view the fine detail of action in meaningful settings has made it possible to examine the phenomena at the core of embodied [9, 14, 34, 53, 54], situated [10, 12, 13, 35, 52] and distributed cognition [31, 15]. The rise of gesture studies in the past decade was made possible by these technological changes and it is now transforming fields such as linguistics [43] and education [20].

New computational algorithms promise to further extend this transformation by enabling automatic recognition, tracking, and summarization of meaningful components of audio-video data [56, 33, 40]. Thus, changes in theory give us new phenomena to see and provide new relevance to things already seen. Developments in digital technology create potential for new tools with which to see those things [11]. These changes and developments are central to HSD goals to understand: human action and development and organizational, cultural, and societal adaptation; how to anticipate complex consequences of change; the dynamics of human and social behavior at all levels, including that of the human mind; and the cognitive and social structures that create and define change.

This proposal takes advantage of these recent developments in digital technologies and their implications for theory in the behavioral sciences. We propose to develop a flexible extensible framework for multiscale analysis of high-density data recordings of complex human activity and evaluate its effectiveness in two critical domains: airline flight instruction, and workstation activity. We will leverage our current collaborations with other laboratories in the areas of digital video (Stanford Diver group), computer vision (UCSD Computer Vision Lab and UCSD Computer Vision and Robotics Lab), and

augmented environments (Rufae Research Network).

The primary motivation for the current proposal derives from our belief that we are at a critical moment in the practice and scope of behavioral research. We argue that the main obstacles to fully capitalizing on this opportunity are the huge time investment required for analysis using current methods and understanding how to coordinate analyses focused on different levels so as to fully profit from the theoretical perspectives of multiple disciplines. The research we propose here is part of an explicit long-term strategy to reduce the cost of performing in-depth analysis, to increase the power of analyses, and to facilitate the sharing of analyses.

A.1.1 Reducing the Cost of Analysis

Today the high labor cost of analyzing rich activity data leads to haphazard and incomplete analyses or, all too commonly, to no analysis at all of much of the data. Even dataset navigation is cumbersome. Data records are chosen for analysis because of recording quality, interesting phenomena, and interaction density—producing a haphazard sampling of the recorded set. Good researchers have a nose for good data, but also have a tendency to focus on small segments of the record that contain "interesting" behavior, analyze them intensively, and then move on to the next project.

When analysis is so costly, few analyses can be done—so datasets are severely underutilized—and researchers come to have a large investment in the chosen data segments. Since each analysis may appear as an isolated case study, it can be difficult to know how common the observed phenomena may be. Larger patterns and contradictory cases can easily go unnoticed. Well-known human confirmation biases can affect the quality of the science when each analysis requires so much effort. Thus, one focus of our proposed research will be on the developing and assembling tools and practices to speed and improve analysis. We will extend our use of computer vision techniques to automatically annotate video data from our focal domains and add facilities to help manage and coordinate both data collection and analysis. The goal is to facilitate the creation, maintenance and manipulation of temporal and sequential relations within and between analyses.

A.1.2 Increasing the Power of Analysis

A significant scientific challenge for all disciplines is how to represent data so as to make important patterns visible. In the behavioral sciences, researchers transcribe and code data in a wide variety of ways, creating new re-representations of the original events [31]. Currently the coordination of multiple re-representations with the original data is typically done by hand, or not at all. Since this re-representation process—including all sorts of transcription, coding system development and implementation, and re-description—is what allows us to do science [21], even small improvements in automating coding, transcription, or coordination of representations can be crucially important. Recent developments in behavioral science theory create special challenges in this regard.

Increasingly theories are concerned with patterns that can emerge from the interactions of many dynamically linked elements. Such interactive patterns may be invisible to approaches that decompose behavior into the more or less independent components created by historical distinctions among behavioral science disciplines. This is why multidisciplinary behavioral science is necessary. But tools that match this multidisciplinary vision are also needed.

Visualizing Multimodal Activity

The richly multimodal nature of real-world human activity makes analysis difficult. A common strategy has been to focus on a single aspect of behavior or a single modality of behavior, and to look for patterns there. However, the causal factors that explain the patterns seen in any one modality may lie in the patterns of other modalities. In fact, recent work suggests that activity unfolds in a complex system of mutual causality. Analysis may still be based on decomposition of the activity, as long as there is a way to put the pieces back together again. That is, as long as there is a way to visualize the relations among the many components of multimodal activity.

Visualizing Macro-Micro Relations and Coordinating Multiple Scales

The structure of the existing academic disciplines attests to the fact that human behavior can be productively described at many levels of integration. Neuroscientists describe regularities at a finer scale than psychologists, who describe phenomena at a finer scale than linguists, who in turn tend to describe behavior at a finer scale than anthropologists. A deep understanding of the nature of human behavior demands not only description on multiple levels, but integration among the descriptions.

As behavior unfolds in time, describable patterns that take place on the scale of milliseconds are located in the context of other describable patterns that display regularities on the scale of seconds. Those patterns in turn are typically embedded in culturally meaningful activities whose structure is described on the scale of minutes or hours. Patterns at larger time scales are created by and form the context for patterns at shorter time scales. Visualizing and reasoning about such nested temporal relations will require representations that allow coordination of analyses across multiple scales.

A.1.3 Facilitating Sharing of Analyses

The high cost of performing analyses on data that represent real-world activity means not only that too few analyses are conducted, but that analyses tend not to be shared. Most often when results of a video analysis are published, neither the activity of doing the analysis, nor the procedure that was used are shared. This creates a situation in which most analyses are idiosyncratic and possibly non-replicable.

The multiscale timeline representation discussed below promise to provide a natural foundation for collaborative analysis. Specialists may do analyses in parallel and then tie them to a shared timeline backbone. Support for the coordination and comparison of multiple analyses may add to the power of the analysis, while simultaneously providing a medium for assessing inter-rater reliability. It will also make the analysis process itself more transparent. Making more of the analysis process visible to a critical audience of peer researchers makes science stronger. Making it more visible to students learning the techniques improves training.

We also expect that facilitating the sharing of analyses will feed back to reducing the cost and increasing the power of analyses as new techniques and analytic practices are discovered and shared.

A.1.4 Research Questions

The following are the motivating research questions we will address:

- Can multiscale representations facilitate identification of relationships among patterns in human behavior described at different levels of integration?
- Can multiscale visualization techniques and lens-based information filtering and management facilities be extended to provide more effective analysis of multimodal data?
- Is it possible using a multiscale timeline to keep the macro context in view while examining a micro analysis? And can such timelines help us address the long-standing problem of visualizing and managing the relationships between analyses performed on differing timescales?
- Can existing computer vision algorithms be used effectively to automatically annotate video data to assist analysis?

Before detailing how we will approach these and related questions, we give a brief example of a timeline-based visualization and then sketch two scenarios to better convey the character of facilities and analyses we are proposing.

In our work on driving, alluded to earlier, we have found timeline-based representations to be invaluable for analyzing activity data (instrument recordings and results of computer vision annotations of video) recorded in an instrumented car. As a simple example, Figure 1 depicts graphs of selected car parameters coordinated by time and linked to GPS-derived freeway locations.

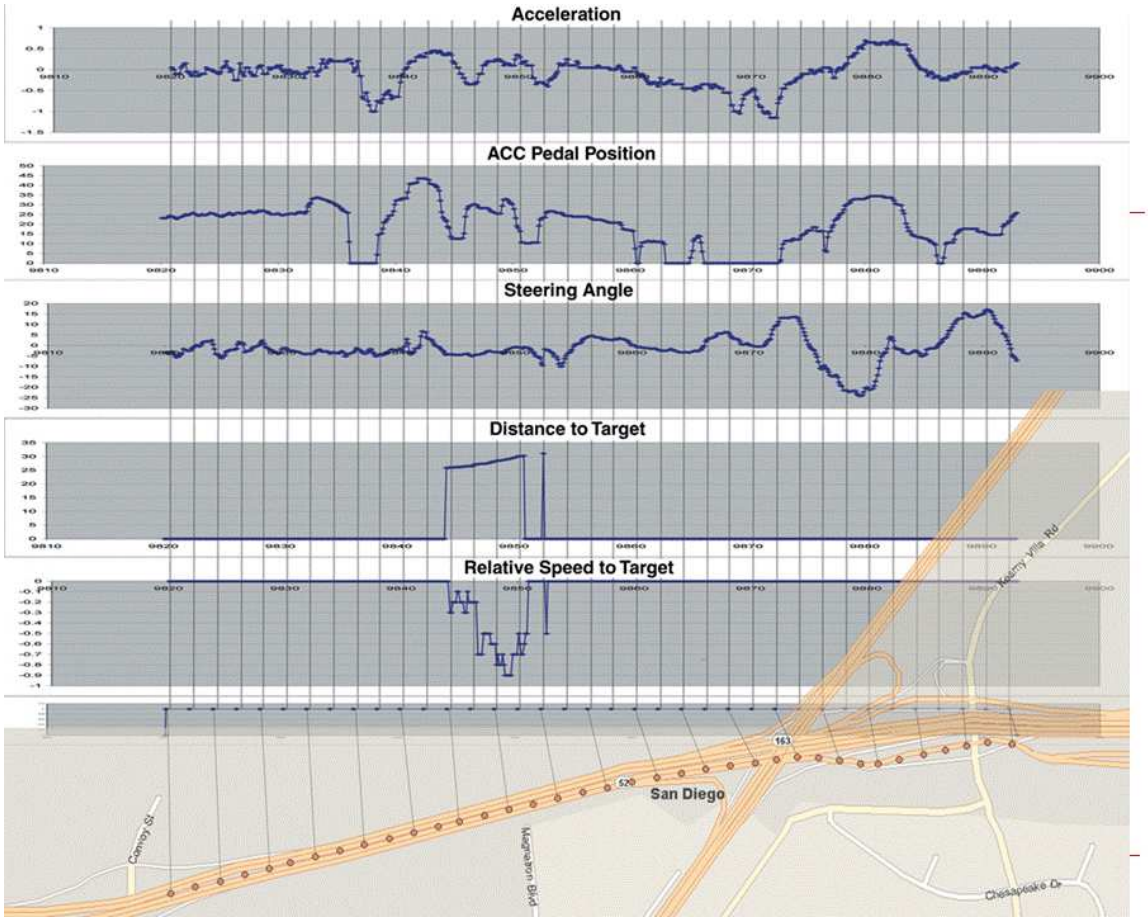

Figure 1: *Results from an analysis tool we developed to allow analysts to graph selected car parameters. This can include results from automated video analyses and can be linked by time or to GPS-derived freeway locations. It is an example of the types of time-based visualizations and linkages we plan to develop in the proposed effort.*

A.1.5 Scenarios

The scenarios presented below illustrate how the tools we envision can accelerate and improve the process of analyzing on-going behavior. Each scenario draws on work that is already underway in our laboratories. The problems described are real ones that have been encountered in this work. The solutions described in the scenarios show how the research we propose could change analysis.

Scenario: Activity in the Airline Flight Deck

In the world of commercial aviation, a Quick Reference Handbook (QRH) is a carefully designed set of checklists and procedures for use in emergencies or abnormal operating conditions. Barbara Tener, a human factors specialist working for a major aircraft manufacturer's operations department, wants to know how pilots from different cultures make use of QRH in the cockpit. She has arranged for a number of airline crews from several regions of the world to fly a challenging mission in a high-fidelity flight simulator. The simulator produces a rich data stream that records all significant flight parameters, instrument readings, and control positions. This data will be useful, but she also needs to know about aspects of the pilots' behavior that are not directly reflected in the behavior of the airplane.

Using multiple digital video cameras installed unobtrusively in the simulator, she has created a rich documentation of the behavior of the

flight crews. In addition to a wide-angle camera recording the entire cockpit, there are special cameras aimed at the pilots and at the slots under the cockpit side windows in which the Captain's and First Officer's QRHs are stowed. Various audio tracks capture the pilots' voices, radio transmissions, and ambient noise.

Barbara would like first to document the variability in the use of the QRH. The behavioral data from each simulated flight were linked during data collection. She synchronizes the data stream from the simulator to the behavioral records. Then, running the analysis off-line, overnight, she creates timelines for each of the simulator sessions using automated object recognition algorithms to code crew retrievals of the QRH from its storage bin. She also instructs the analysis program to mark the onset times for each abnormal condition as recorded in the synchronized simulator data stream. These times are used by the analysis program to re-enter the video and audio data and collect a set of summary videos, each one capturing events from just before the abnormal condition arose to the end of the crews' first action in response to the condition. When she returns to work in the morning, each simulator session run has been re-represented as two timelines (one for abnormal conditions, and one for QRH retrievals) and a collection of summary videos (each one capturing a proposed abnormal condition, QRH retrieval pair).

Because things can get chaotic in the cockpit when emergencies occur, Barbara wants to review the set of videoss collected by the program, making sure that each observed use of the QRH is matched to the correct triggering condition. That is an expert judgment, so Barbara asks Chuck, one of the company engineering test pilots to join her. Together, Barbara and Chuck review the videos detecting and repairing a few mistaken classifications. With the condition/QRH retrieval pairs coded, they create a composite timeline showing the linked pairs of events.

The composite timeline is a powerful representation. Because temporal relations are represented as spatial relations in the timeline, Barbara and Chuck can do conceptual work using perceptual processes. They can get a feel for the organization of the data, for example inferring that crews respond more quickly as the experiment goes on, by visual inspection. The composite timeline also contains a measure of the latency between the first indication of the failure and the crew retrieving the QRH. The latency measure can be exported from the timeline to an off-the-shelf spreadsheet. Chuck is surprised to see such long latencies in some cases, and goes back to the video data one more time to see what is accounting for this. He decides he is simply seeing the difference between test pilots and airline pilots.

A number of tests are now possible with respect to QRH mobilization latency. If the within-flight variability is greater than the between-flight variability, it suggests that the properties of the setting are the controlling variables. Inspection of the timelines permits Chuck and Barbara to return to the video data with a new set of questions about the design of the cockpit, the QRH, and the airplane procedures. Using the multiscale interface to manage the inventory of data records Barbara collects the records by culture. Is the within-group variability greater than between-group variability? If so, that suggests that individual differences may swamp the effects of culture.

Documenting the variability in QRH usage as described above would be prohibitively expensive using the tools and methods available today. And that is just the first step in a much deeper study of the dynamics of human behavior in this important work setting. These quantitative results allow researchers to make principled decisions about which specific instances merit closer scrutiny. Researchers can return to the video data knowing whether the events chosen for qualitative analysis are typical or rare.

This scenario illustrates how analysis can be jump-started by the automatic creation of timelines. It also illustrates how these representations can facilitate collaborative work, the checking of automated codings, noticing and testing

quantitative relationships, the flexible management of multiple data records, and improved information about the nature of the sample of events chosen for analysis.

We already have in place an agreement with a major airframe manufacturer that will give us access to data of the sort described here. Co-PI Hutchins has eighteen years of experience studying the cognitive consequences of cockpit automation. The role of culture in cockpit operations has recently become a critical issue in the commercial aviation industry, but empirical studies based on careful analysis of flight crew behavior are just beginning [30, 46]. This will be one important focus of our proposed research. A description is provided below.

Scenario: Adaptive Response of a Ship's Navigation Team

Three researchers are working together to construct a detailed analysis of video recordings of the activities of the navigation crew on the bridge of a ship. They have created a macro-level timeline of the changing configurations of the team's activity before, during, and after the failure of an important piece of navigational equipment. This timeline documents the adaptive response of a social group to a perturbation in their working environment. Descriptions at this level of analysis clearly show the dynamic response: the quality of the output of the navigation team declines sharply when the equipment fails, but then recovers in fits and starts over a period of about 40 minutes. Changes happen in the division of labor among the members of the team, in the way tools are used, in the way language is used, and in the organization of the computation the team is performing. Thus, at a relatively macro level, it is possible to describe social change as interlocking dynamic reconfigurations of several aspects of the activity in multiple modalities.

But what lower-level mechanisms make this adaptive response possible? What role do the moment-to-moment dynamics of social interaction play in the reconfiguration of the division of labor among the members of the navigation team? The researchers can show that the introduction of a new tool changes the relations of the navigators to each other and to the activity, and that seems to facilitate explicit reflection on the organization of the computation they are performing. But how do the changes in the different sub-systems interact to produce the observed outcomes? Stated in this way, this is a classic example of the problem of finding the relations among descriptions rendered at different levels of integration of events. The long-standing issue of the relationship of macro to micro descriptions and processes is an instance of this problem.

One of the timeline representations encodes the quality of the series of navigation fixes performed by the team (bottom panel in Figure 2). This is one of many representations of the macro-level processes. Using representations at this level, the researchers have identified what they consider to be a set of inflection points in the adaptation of the team to the equipment failure. These are points where the working configuration of the team seems to undergo dramatic change.

Moira uses the macro timeline to enter the original video data stream at the first identified inflection point. Her work here consists of taking an inventory of sources of information that enter into the computation performed by the navigation team. She documents when bits of information become available in the environment and how they are incorporated into the computation process. She does this by creating a new timeline on which she locates the first appearance of each element that is subsequently incorporated in the computation.

Ted examines the original timeline. It indicates the time at which each computational step was performed. By juxtaposing Moira's new timeline with the original timeline, he can see (as a visual pattern) that early in the adaptive response the order in which the navigation team incorporated new elements in their computation was almost entirely driven by the order in which the terms appeared in their working environment. This means that the sequential organization of the navigation team's behavior was probably driven by aspects of the working envi-

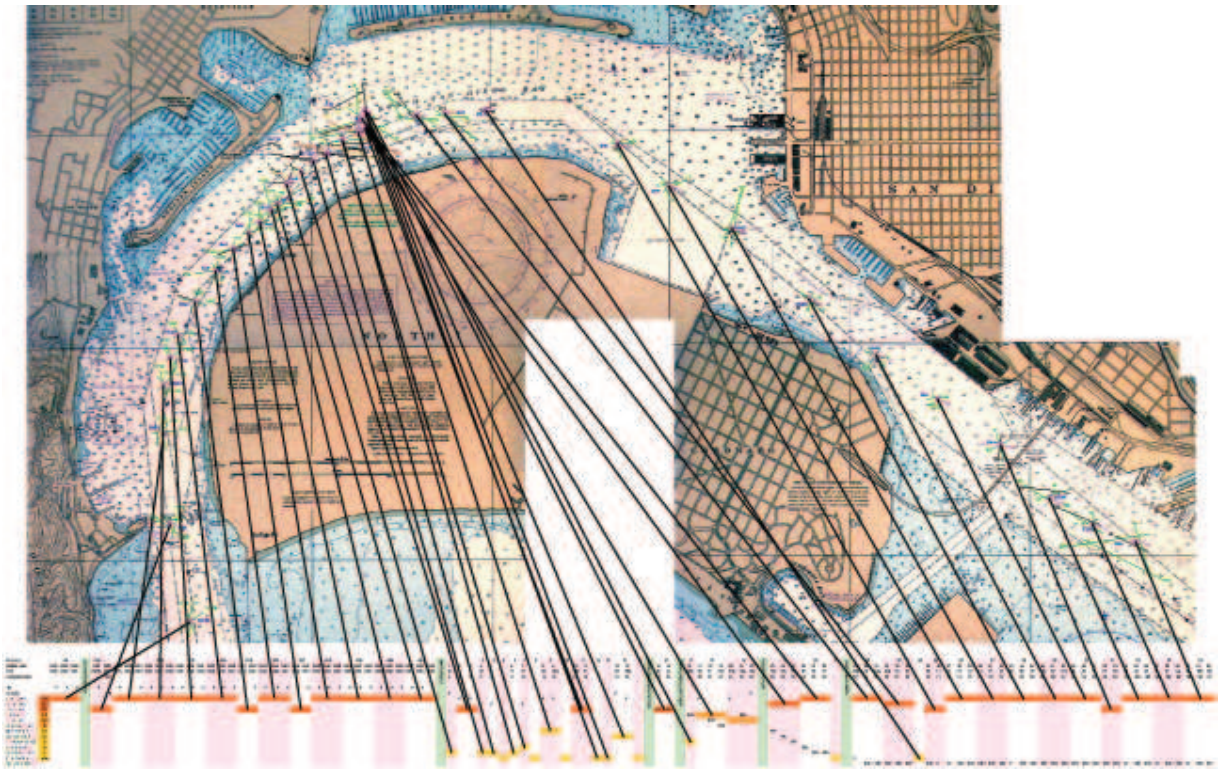

Figure 2: *Analysis timeline linked with ship position.*

ronment rather than by an internal plan.

Ted and Alice then examine the original video at the inflection points looking to see how the navigator's body and the tools interact. Moving the digital video one frame at a time, they notice that at a key inflection point the navigation plotter moves his plotting tool on the chart in a way that would (if it could be justified) improve the quality of the position fixes. Using his body and his tools in coordination with the chart, the plotter seems to imagine a hypothetical world in which the position fixes are improved. A moment later the plotter declares, "I know what it's doing!" He then follows through by introducing a new term to the computational procedure that does in fact improve the quality of the fixes. Ted and Alice use the multiscale tools to create a single representation in which the movement of the plotting tool (millisecond timescale) is embedded in the creation and examination of an hypothetical fix triangle (seconds timescale), which is embedded in the discovery of the new term (minutes timescale), which is located at a particular point in the adaptation event.

This scenario illustrates how the proposed work might enable flexible coordination of macro and micro level representations of the activities of a multiperson team jointly engaged in a complex task. We hope it helps make the potential analytic value of the facilities clear. The timeline representations described in the scenario would make possible powerful new kinds of analysis. However, at present the construction and coordination of timeline representations is prohibitively expensive. The coordinated timelines shown in Figure 2, for example, required several person-months of effort to construct. Members of our laboratory are collecting digital video of multiperson activity in a number of domains (e.g., dental hygienist training, everyday activity in Japanese homes, first-responders to emergency situations, scientists working in their laboratories) and in all of them there are exciting opportunities to employ the proposed facilitates to better understand activity dynamics across time scales.

A.2 Proposed Research

In the long term we envision a widely shared research infrastructure to advance understandings of the dynamics of human activity by supporting the types of analyses characterized in the scenarios above. The goal of the work we propose here is to iteratively develop an initial analysis framework and evaluate it in a set of projects carefully chosen to demonstrate the potential of automatic annotation and multiscale visualization. To move towards this goal we will:

- develop a multiscale timeline for annotating, visualizing, and navigating video and other rich time-based data
- identify existing computer vision techniques that promise to provide effective general support for automatic annotation of video records and explore how to make them available to analysts while requiring minimal computer vision expertise on their part
- evaluate the developing framework to (1) assist analysis of pilots' behavior in flight simulators, (2) study and characterize workstation activity, and (3) represent the temporal structure of the writing process
- make the framework available to members of our laboratory and collaborators[1] (UCSD colleagues Morana Alac, Mike Cole, Chris Halter, and Randy Souviney, Louis and Kim Gomez at Northwestern and UIC, Chuck Goodwin at UCLA, and Roy Pea at Stanford) in order to provide us with the feedback necessary for iterative development
- compare automatic and manual video summarization as a basis for helping people to return to previous work contexts

Our primary concern is to better understand and support the practices and workflow involved in analyzing actual instances of human activity (generally in naturally occurring cultural contexts) and in interpreting, within the wider contexts in which they are situated, the cognitive and social import of the fine-scale details of the activity. For the work we propose this has both a methodological and an meta-methodological aspect.

All methods for analyzing activity involve creating a cascade of representations down stream from the original data records. In the case of video this means multiple transcriptions (e.g., of verbal behavior, bodily behavior, and interactions with material and social resources), followed by the application of coding schemes, and the production of graphical representations of relations among the entities identified in transcription and coding. One of our goals in this work is to automate the creation of some of the first representations in this cascade. By inserting automatically generated timelines that are populated by events and objects identified by computer vision techniques, we hope to increase the volume, quality, and ease navigating and analyzing the cascade of representations that follows.

The meta-methodological aspect derives from the fact that we are always examining our own practices and attempting to find more powerful tools or more appropriate techniques for doing our work. We are as interested in the so-called soft technologies of research practices as we are in the hard technologies of digital media and computation.

A.2.1 Research Foundations

Because the proposed research has grown out of our recent work and collaborations we first discuss these efforts. We end each section with a highlighted description of how the effort has influenced our plan for the proposed work.

Dynapad

Dynapad[2] is a multiscale interface and information visualization environment developed at UCSD [1, 2, 25]. It is the third generation in the Pad++ lineage of zoomable multiscale systems [5, 4]. The goal of the effort is to understand

[1] The support letters from Cole and Souviney are representative of the excitement our colleagues have for the proposed work and of their willingness to collaborate.

[2] The development of Dynapad was supported by an NSF grant (ITR-001389).

Figure 3: *Sample PDF portraits of a paper describing and its references.*

the cognitive strategies people use in managing information collections in visual workspaces and how to design multiscale representations and a versatile infrastructure of tools to support those strategies. Dynapad currently supports multiscale organization and management of collections of digital photos and iconic representations of journal articles.

Figure 3 depicts a collection of iconic representations of PDF files. This collection includes a recent paper about Dynapad in the center. The icons surrounding it are representations of the papers it references. We find these automatically constructed image-based iconizations (the first page of a paper with a montage of images from the paper arrayed on it) to be effective reminders of a paper's content. Clicking on an icon loads the PDF in a viewer for reading. We have also explored the usefulness of informal piles to support sense making and structuring of collections of journal articles and of images from digital cameras[1, 2].

Figure 4 depicts a portion of a Dynapad workspace showing informally structured piles of papers. A timeline lens is positioned over piles of documents to temporarily array them along a time continuum. In this case the dates stored inside the PDF files at the time of their creation are used to layout the members of the pile. The lens can also exploit the file date to array the same set of papers in terms of when they entered the file system. The former might be useful for judging the lineage of the papers while the latter allows easy access to a particular paper based on when one remembers obtaining it. We expect informal spatial structurings and the dynamic lens architecture we have developed for Dynapad to be particularly useful for supporting multiscale timelines and management of multiscale analyses of time-based activity data.

In the proposed effort we will develop Dynapad lenses for examining video data, provide flexible pile-based facilities for structuring time-based data, and examine multiscale representations of video and other time-based data. Our plan is to explore techniques similar to those we developed for multiscale piles. For example, as one zooms out of the piles the more representative images become larger. We expect that similar approaches will be effective with video. We will also draw on the growing video summarization literation. See [36] for a review.

Simile Timeline

The Simile (Semantic Interoperability of Metadata and Information in unLike Environments) project (`http://simile.mit.edu`) at MIT recently made available a very interesting open-source timeline facility. Simile Timeline is a DHTML-based widget for visualizing time-based events in web browsers. It is similar to Google Maps in that it can be used without a need for software installation and a timeline can be populated via an XML data file.

A Simile Timeline contains one or more bands that can be panned by mouse dragging. A band can be configured to synchronize with another band such that panning one band also pans the other in appropriate ways, even when the two bands represent different time scales. This is a mechanism that can provide a limited form of multiscale access since each band can control the mapping between pixel coordinates and dates/times.

We will integrate Dynapad with Simile Timelines to support multiscale analysis and navigation of time-based data. There are numerous tradeoffs between web-based interfaces and locally running applications, especially for large video data corpora. We will explore this tradeoff space so as to be able to provide analysts with an effective environment that we expect will incorporate

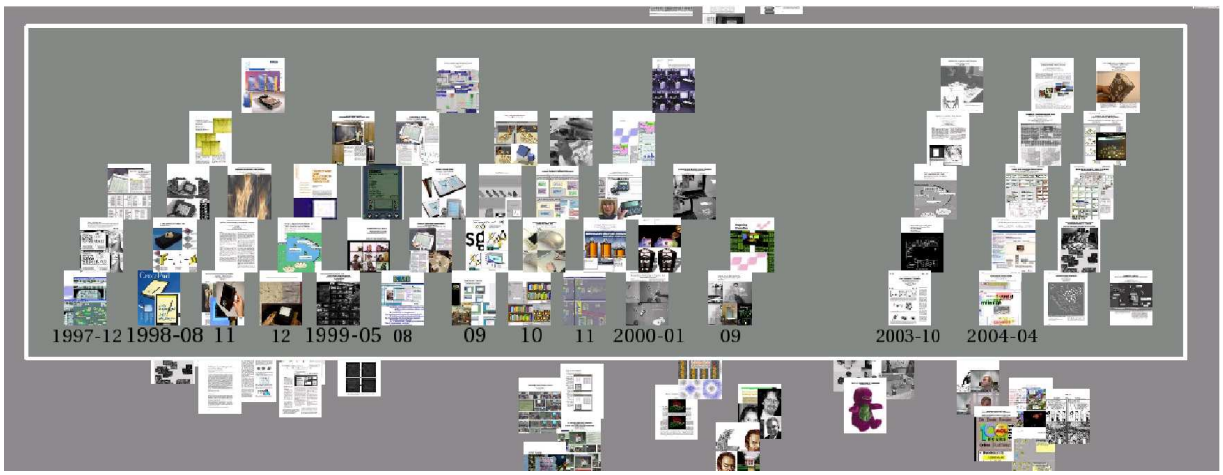

Figure 4: *A timeline lens is positioned over document piles in a Dynapad workspace. The lens provides a temporary chronological ordering (by either PDF-creation date or date of entry into the file system) of the documents. Year and month are indicated along the bottom of the lens.*

both.

The Simile project has also recently added a lightweight structured data publishing framework (Exhibit) that doesn't require a database but provides database-like facilities (e.g., sorting and filtering) within a web page. As with the integration of Dynapad and Simile Timelines we will examine how to best use a combination of web-based and local facilities to support analysis and collaboration.

DIVER

Diver (for 'Digital Interactive Video Exploration and Reflection') [48] is an application developed at Stanford for repurposing and commenting upon audio-video source material. We have recently initiated a collaboration with Roy Pea and other designers of Diver. Access to Diver software and support from the Diver team will be extremely useful for the work we propose here.

The Stanford project and our colleagues in the Teacher Education Program here at UCSD are focusing on use of Diver to record activities in classroom settings using digital video cameras. The current Diver user interface is shown in Figure 5. Users can zoom, pan, and tilt a virtual camera (indicated by the yellow rectangle in Figure 5) in the source video window to dynamically record a video segment. A *Dive* involves creating a collection of clips arrayed in panels that show thumbnails and allow textual annotations. The clips are independently playable. A web-based version of Diver, WebDiver, is also available. WebDiver can be used to upload the resulting clips to the web to allow shared access and enable collaborative commentary.

For our purposes, Diver provides an ideal tool for extracting videos clips in the service of analysis, reflection, and annotation. Being able to easily and precisely refer to video source segments and annotate them, previously available only through general-purpose video editing software, makes video potentially as plastic and portable as text. To be able, while the video is running, to move the virtual camera around and change its size and thus the portion of the video frame recorded (e.g., the hand gestures or facial expressions of one the participants) provides a simple natural way to focus analysis.

We will integrate Diver and WebDiver into our multiscale analysis framework to support extraction of video clips, their annotation, and the ability to link them to multiscale timeline representations. We will also support linking between Diver clips, timeline representations, and spreadsheets. We have found that spreadsheets are commonly used to code and annotate video data and we want to understand, exploit, and

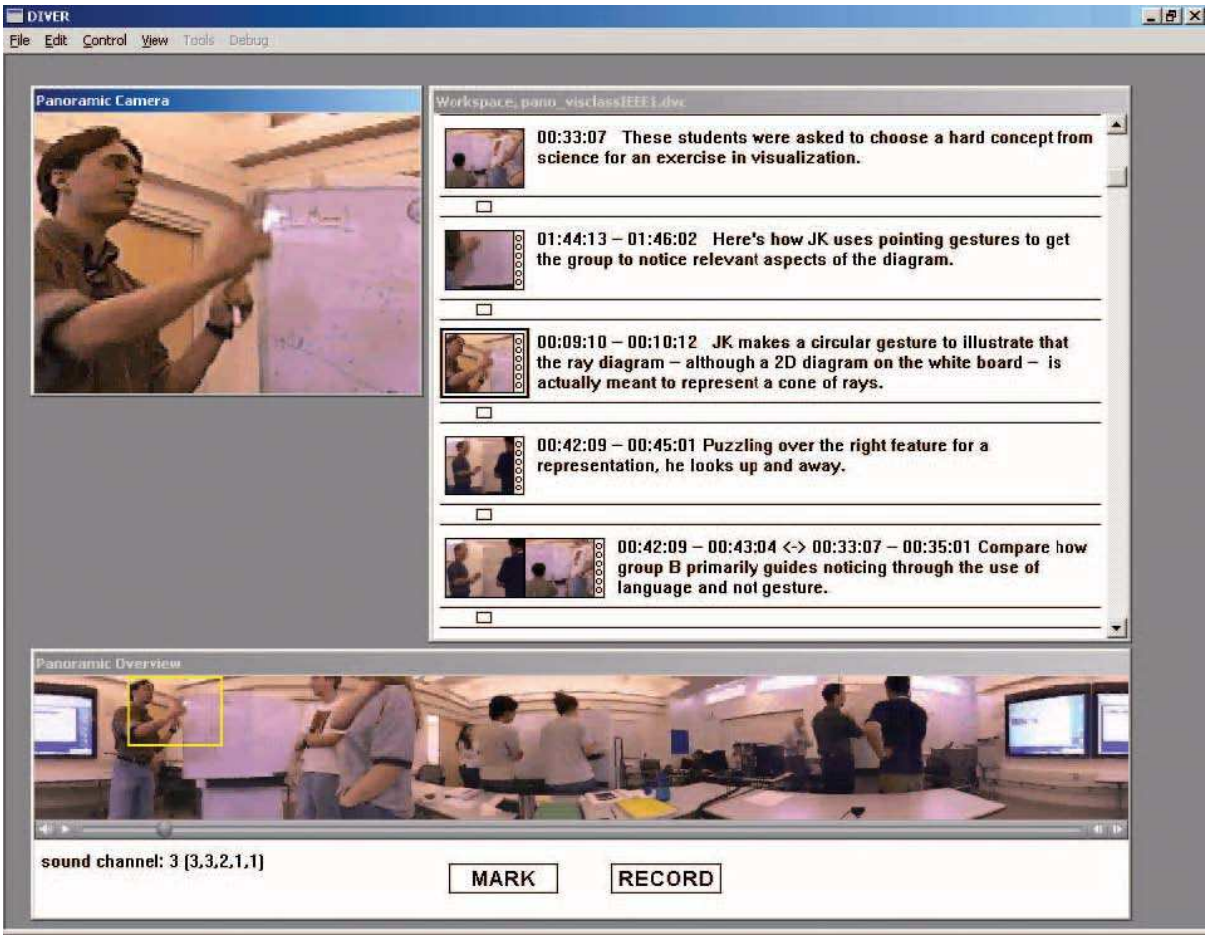

Figure 5: *DIVER (Digital Interactive Video Exploration and Reflection). The yellow rectangle on the left allows positioning of the virtual camera with the mouse to frame sections of interest. While the video is running the virtual camera can be positioned and zoomed to focus on areas of interest. The mark button allows specific positions in the video to be marked for return or replay. Pressing the record button begins recording of a clip. During recording the virtual camera can be panned and zoomed. Annotated clips are shown on the right. These and associated videos can be published directly to the web. See http://diver.stanford.edu for more information.*

support that practice.

To effectively link video clips into our timeline-based analysis framework will necessitate abstracted graphical representations of the clips. There are numerous approaches for abstracting and summarizing videos. These range from simple sampling-based keyframe extraction to more sophisticated techniques that attempt to analyze the semantic content of videos. See [36] for an overview and review. While we will draw on this literature and additional computer-vision techniques discussed below, our focus will be on developing simple and general representation and annotation facilities. We expect keyframing and multiscale timelines to be extremely useful. We will support links between video and analysis bands so as to be able to flexibly move between them. We will also explore overlays on the video itself to represent how segments of video data have been coded drawing on modern compositing techniques [8]. In addition, analysts will be able to easily zoom out and see all the frames of a video segment with overlayed codes. We expect these and related views to serve as natural indexing mechanisms for navigation of data and analyses.

SIDGrid: The Social Informatics Data (SID) Grid

SIDGrid is an NSF-supported computing infrastructure being designed to provide integrated computational resources for researchers collecting real-time multimodal data at multiple time scales. Data is stored in a distributed data warehouse that employs Web and Grid services to support storage and access.

The main tool in the existing SIDGrid infrastructure is Elan. ELAN (EUDICO Linguistic Annotator) is an open-source annotation tool that allows one to create, edit, visualize and search annotations for video and audio data. It was developed at the Max Planck Institute for Psycholinguistics, with the aim to provide support for annotation of multi-media recordings. ELAN is specifically designed for the analysis of language, sign language, and gesture. Figure 6 shows the main Elan screen. Members of our laboratory have experience with Elan.

We will explore use of ELAN as a component of our analysis framework and compare and contrast it with Diver. More importantly we will ensure that our overall effort can be integrated within SIDGrid and connected to the associated web and grid services. One of us (PI Hollan) presented at the December 2006 SIDGrid meeting in Chicago and it is clear that we be welcomed as members of this new community. In addition we should mention that SIDGrid is part of the TeraGrid open scientific infrastructure and UCSD is one of the nine partner sites on the TeraGrid.

Computer Vision

Computer vision techniques have advanced in capabilities and reliability to the point that they promise to be highly useful tools for aiding analysis of video data. To characterize this potential we first describe our recent experience [7] developing techniques to automatically annotate video of driving activities to assist a research team in understanding the cognitive ecology of driving and in designing instrumentation and controls to improve driver safety. In order to ground design in real driving behavior we instrumented an automobile to record multiple video streams and time-stamped readings of instruments and controls. This included video from a head-band *3rd-Eye* camera we developed (Figure 7) and from an array of ten cameras positioned to capture views from within and around the car, as well as of the driver's face and feet.

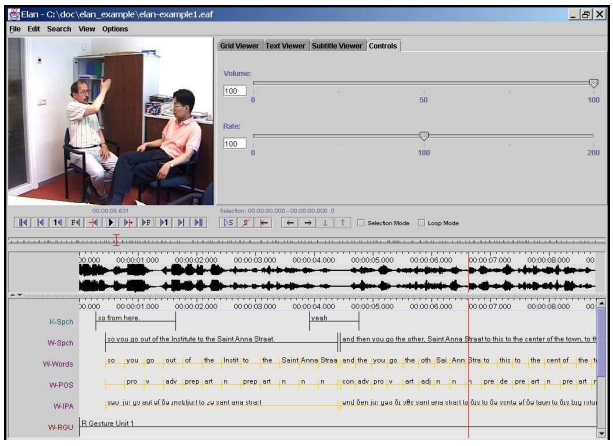

Figure 6: *The main Elan screen has a video viewer, annotation density viewer, waveform viewer, timeline viewer, and media player controls.*

Figure 7: *Top: Head-Band "3rd Eye" Camera (left and middle) and view from it (right). Bottom: Example views from cameras. (Composite on left, Omniview in center, and Rear view on left)*

We are extremely encouraged by our success in automatically annotating video to assist analysis. For example, we developed simple matlab code to compute the lateral angular velocity of the head from the 3^{rd} Eye camera video. This allowed identification of even small head position adjustments as well as glances to the rear view mirror, glances to the left or right side mirror, and large over-the-shoulder head movements. We also thresholded the amplitude of

Figure 8: *SIFT object recognition system[39]. From two training images (left), invariant features are automatically extracted. When presented with a new, cluttered image that has extensive occlusion (center), the system uses the previously extracted features to recognize the objects (right). A box is drawn around each recognized object. Smaller squares indicate the locations of the features that were used for recognition. Figure from [39].*

recorded audio to index times when someone was speaking in the car. Foot motion and lateral foot position were extracted from a "FootCam" video using a simple detection algorithm. In combination with recordings of brake pedal pressure this enabled easily determining, for example, when drivers move their foot to the brake pedal in preparation for braking. We also developed code to determine where the hands were positioned on the steering wheel and to automatically compute lateral position of the car as a basis for detecting lane changes. Our experience applying these techniques to video data [42] will be invaluable for the research we propose here.

While we do not have space to review all the vision-based techniques we see as applicable and promising for automatic annotation, we briefly mention three examples: object recognition, face and emotion detection, and pose estimation. It is important to again note that our focus is not on developing algorithms but evaluating existing algorithms for use in our analysis framework. In addition, unlike most work in computer vision, we do not typically require real-time processing. For our purposes offline processing is usually sufficient.

Object Recognition It would be a boon to digital video analysis if the computer could automatically label all (or even most) frames or segments of video in which a particular object is present. For example, if analysts are interested in activities involving interaction with specific objects, they might want to view only those segments of video that involve those objects. To evaluate this capability, we will explore a number of recognition methods. One very promising candidate uses distinctive invariant features extracted from training images as a basis for matching between different images of an object. An important aspect of the SIFT (Scale Invariant Feature Transform) technique [39] is that it generates large numbers of features that densely cover the image over the full range of scales and locations. The features are invariant to image scale and rotation, and provide robust matching across a substantial range of affine distortion, addition of noise, change in viewpoint, and change in illumination. For example, objects, such as the toy frog and train depicted in Figure 8, can be found in complex scenes with extensive occlusions.

Since this algorithm is probabilistic, we can allow the user to modify the algorithm's threshold depending upon the task. For example, the threshold for object detection could be set at a low value in which virtually every frame that contains the object is detected, with the price of having some false alarms (flagged frames in

Figure 9: **Left:** *The 3D tracking algorithm GFlow at work in outdoor conditions. The points on the face are not physical, but are rendered by the computer to demonstrate that it has tracked the desired facial features. The algorithm simultaneously estimates the pose of the head (3 dimensional reference axis in bottom left corner) and face deformations. Once the head pose is known, multiple camera angles* **(Center)** *can be synthetically rotated and merged to render the face from any desired viewpoint* **(Right)**.

which the object is not actually present). In this case, a small amount of user intervention would be required in order to cull the false alarms from the true detections. On the other hand, the object detection threshold could be set at a high value in which there would be virtually no false alarms (every flagged frame is a true detection), with the price that in some frames the object would be present but not detected. Depending upon the analysis task (finding every instance vs. finding a collection of representative instances), one or the other threshold (or somewhere in between) might be appropriate.

Face and Emotion Detection There are myriad ways in which computerized face detection and face tracking could enable new types of analyses with huge potential gain and minimal time commitment on the part of the analyst. Current face detection algorithms [55] could be employed to annotate the video so that appropriate video segments could be located quickly and accurately. One example of the state of the art in current research is work by one of our recent Ph.D. students[3] [41] to automatically annotate a video with the subjects' emotions, as determined by their facial expressions. Computerized facial expression analysis can be done with existing technology on frontal views of faces [38]. To analyze facial expressions from non-frontal views of a person's face, sophisticated 3D tracking algorithms such as G-flow [41] can be used to find the 3D pose of the face and the 3D locations of key points on the face from 2D video of the subject. By fitting the 3D locations of these key points to a database of laser scans of human heads [6], we can synthetically rotate the face from any viewpoint to a frontal view (see Figure 9), from which the emotion of the subject can be determined using the aforementioned facial expression analysis system.

Head Pose Estimation Erik Murphy-Chutorian just defended his dissertation proposal (PI Hollan is a member of his committee). Erik plans to investigate pose-invariant head and hand detection, head pose estimation (see Figure 10), tracking of heads and hands across multiple views, head and hand gesture extraction, and spatio-temporal analysis of human activities.

We will select computer-vision techniques that are promising for use in the types of automatic

[3]The graduate experience of Tim Marks exemplifies the type of cross laboratory training we expect to provide for graduate students associated with the proposed project. While PI Hollan was the chair of his committee, Marks worked both in the Distributed Cognition and HCI Lab in Cognitive Science and the Machine Perception Laboratory in the Institute for Neural Computation, where he was supervised by PI Movellan. Currently Tim is a postdoc in the Department of Computer Science and Engineering.

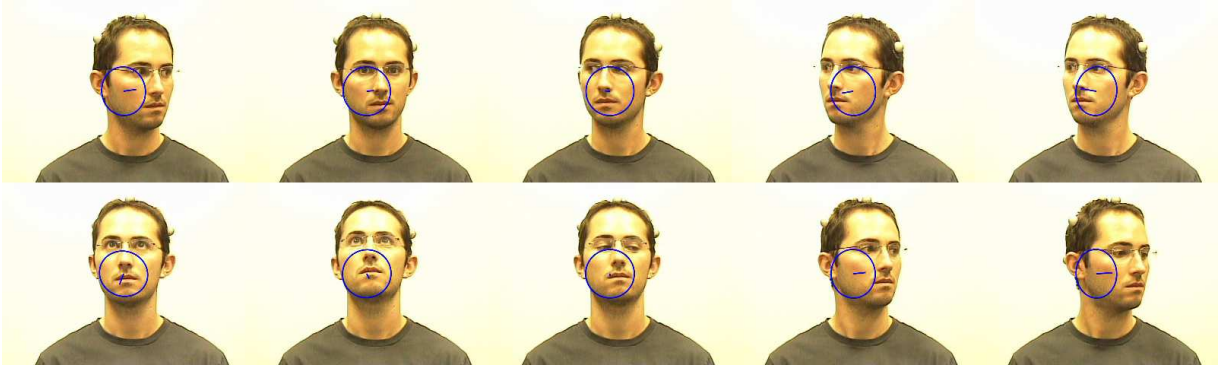

Figure 10: *Example pose estimates smoothed over time. SIFT-prototype detectors were applied to every frame and the pose track was smoothed using a Kalman filter. Pose is indicated by the direction of the blue circle overlay.*

annotation we described above. They will range from techniques for simple recognition of objects to facial, emotion, and pose orientation detection.

In addition, we will explore interfaces to allow analysts to parameterize these facilities and connect them together using a visual interface much like the recent Yahoo Pipes interface (`http://pipes.yahoo.com`). For our work the feeds will be video streams (rather than RSS feeds) with connections to recognition and visualization facilities. Our goal is to ease connecting and parameterizing annotation functions in order to make them more accessible to analysts.

While we expect that our students and collaborators will explore use of our developing analysis framework in a wide range of areas, we propose to focus our efforts on three areas: understanding how airline pilots learn in flight simulators, normal workstation activity, and the writing process. The first area is chosen because the activity is dynamic and richly multimodal and the outcomes have social significance. In addition, the fixed nature of the simulator environment will allow us to exploit that structure to simplify vision-based annotation in ways similar to recording in automobiles. The second area was selected because of the increasing importance of workstation activity in almost every aspect of professional and personal life. Here again we expect vision-based annotation to benefit from the known structure of the display and interface components as well as methods of interaction. In addition, as in our work studying driver activity in which we had access to all instrumentation, with the workstation we have detailed access to the process and applications that are running and in some cases even to video of workstation users. Finally, because of developments in another research project we have easy access to capture and record the content and temporal structure of paper-based writing activities. We will capitalize on our recent experiences with digital pen software to provide us with a different form of time-based data, the analysis of which should help us to ensure a greater generality of our framework.

In the next sections we discuss these three domains and a novel application of repurposed video from workstation recording to help people to reestablish the context of earlier activity.

A.2.2 Flight Simulator Activity

Under a multi-year agreement with the Boeing Commercial Airplane Group, co-PI Hutchins has negotiated access to training activities at a number of airlines outside the US. (Security provisions enacted in the wake of the 9/11 terrorist attacks have made work with US airlines impossible.) Boeing's interests lie in specific applications and interventions concerning training, operating procedures, and flight deck design in the next generation of airline flight decks. Boeing's goals are not the focus of the work we describe in this proposal.

However, the video data generated by the Boeing project are available to us for other sorts of analysis. In addition, we already have a Human Research Protections Program protocol in place for collecting and analyzing this data. (UCSD HRPP Project No. 050582, "Flight Deck Culture for the Boeing 787")

Hutchins has already collected video data in Japan and New Zealand, and expects to acquire additional data from Australia and Mexico in 2007. In addition to collecting data in simulators located at the training centers of non-US airlines, the project also collects data in simulators located at the Alteon/Boeing training center in Seattle.

These data provide an absolutely unique look at complex, highly structured, expert activity in a setting that is spatially, temporally and institutionally constrained. These properties make this data ideal for the purposes of the current proposal. The activities that we record in high-fidelity flight simulators are complex. They involve the production of multimodal acts of meaning making that are embedded in social and material context. The script-like structure of phases of flight and of flight deck procedures provides a common framework with respect to which the activities of different pilots and even different populations of pilots can be compared. This attribute of the activity also makes it a good early choice for exploration with time line representations because there are recognizable shifts in activity structure in successive phases of flight. By the time they are in training for commercial air operations, pilots have high levels of expertise. Studying expert real-world skills is important, but difficult to do because analysts must have considerable expertise themselves in order to interpret the significance of the presence (or absence) of particular behaviors. Fortunately, co-PI Hutchins' years of experince as a jet-rated pilot and as an aviation researcher provide the necessary analytic expertise. The spatial and temporal constraints on activity in the flight deck make data collection tractible in the sense that recording equipment can be installed in fixed locations, or attached to the clothing of the participants, and the activities to be recorded are sure to take place in an anticipated amount of time (usually about two hours). The institutional constraints guarantee that the data recordings will be rich in observable activity (little down time) because of the high cost of operating the simulator. Finally, it is sometimes possible to acquire a rich digital data stream from the simulator itself (this depends on the practices of the training department involved). Time synching simulator data to the observational data provides a documentary richness that is simply not possible in most activities. It is, however, an exact mirror image of the data collection strategies to be pursued in our proposed investigation of workstation activity described in the next section.

Our observations in Japan have already revealed that language practices in the Japanese airline flight deck can be seen as adaptations to a complex mix of exogenous constraints. Institutions, such as regulatory agencies, adapt to the constraints of global operations when setting the rules that govern airline operations; the decision that air traffic control communications shall be conducted in English, for example. Airlines adapt to the regulatory environment, the marketplace, the characteristics of their workforce and the nature of the technology when setting training and operational policies. Pilots adapt to the residues of the adaptive behaviors of institutions in constructing meaningful courses of action in flight [30].

Commercial aviation is a complex socio-technical system, that has developed most rapidly in North American and Europe. Because we are working with non-US airlines, we are also able to examine how other cultures integrate the practices of commercial aviation into their particular cultural and cognitive ecology [46]. This is a unique perspective on the globalization of one of the most complex socio-technical systems in today's world. Dramatic changes in the demographics of the global population of commercial airline pilots are currently underway. For example, the mean age of pilots worldwide is rapidly decreasing as aviation expands in Asia and Latin America.

The data collected in the Alteon/Boeing training center is expecially interesting from a HSD point of view because it involves American flight instructors working with pilots from other nations. These data permit us to examine the dynamics of a special case of intercultural learning. Airline pilots everywhere share certain elements of professional culture, but in intercultural training, professional culture becomes a resource for overcoming the boundaries of national culture. They permit us to see the contextual grounding of intercultural communication and learning [30].

Pilot and instructor behavior in flight simulator sessions is a class of dymanic human activity that is both theoretically interesting and methodologically tractible. The intercultural aspects of the globalization of commercial aviation add to the societal significance of this work. Our attempts to analyse this class of activity also provide an excellent testbed for the development of tools that can enhance our understanding of the dynamics of human activity.

A.2.3 Workstation Activity

There is a long history and a recent resurgence of interest in recording personal activity. Personal storage of all one's media throughout a lifetime has been desired and discussed since at least 1945, when Vannevar Bush published *As We May Think*, positing the Memex, a device in which an individual stores all their books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. His vision was astonishingly broad for the time, including full-text search, annotations, hyperlinks, virtually unlimited storage and even stereo cameras mounted on eyeglasses.

Today, storage, sensor, and computing technology have progressed to the point of making a Memex-like device feasible and even affordable. Indeed, we can now look beyond Memex at new possibilities. In particular, while media capture has typically been sparse throughout a lifetime, we can now consider continuous archival and retrieval of all media relating to personal experiences. For example, the MyLifeBits [19] project at Microsoft Research is recording a lifetime store of information about the life of Gordon Bell. This includes not only video but the capture of a lifetime's worth of articles, letters, photos, and presentations as well as phone calls, emails, and other activities. This and related projects are documented in the recent series of ACM CARPE workshops on capture, archiving, and retrieval of personal experiences.

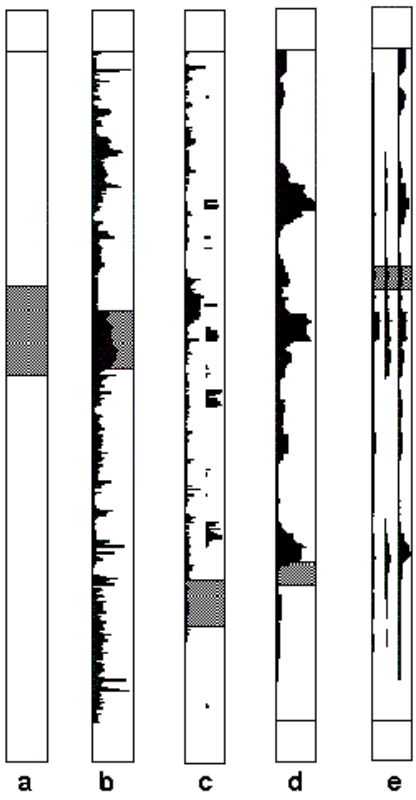

Figure 11: *Examples of attributed-mapped scroll bars from our early work on ReadWear and EditWear. One the left is a normal scroll bar and on the right are examples of visualizing the history of editing activity. We were able to show data such as who edited and the time taken in editing.*

Although what we propose is related to and encouraged by this general zeitgeist, our proposed effort is primarily complementary to other work in this area. Instead of focusing on the technology to collect, store, and access rich video and other records of activity, we will focus on timeline-based multiscale visualization and anal-

ysis of workstation and writing activity.

History-Enriched Digital Objects

We have long been interested in visualizing activity histories. In our early work on ReadWear and EditWear [26] we modified an editor to collect detailed histories of people editing text or code and made those histories available in the scrollbar of the editor (see Figure 11) in ways to inform subsequent activity.

Over the last few years we have conduced a series of pilot projects in which we collected workstation activity of users. While there are commercial software products [4] to record such activity, none of these products can be modified in ways needed to be integrated into experimental analysis frameworks or linked with our multiscale visualization facilities. Like other researchers we have been tempted to build our own software to perform such recording functions. For example, in one of our efforts we built facilities to record low-level operating system call activity on a workstation and then explored parsing that low-level activity record into higher level activity descriptions. The motivation was to have recording facilities that did not require modification to any applications our participants used in their normal workstation activity.

We are now convinced that unless recording and analysis software is openly available and integrated within a shared infrastructure it is unlikely to be useful for the growing research community interested in understanding the dynamics of human activity. In our proposed work we will make use of the recently available Glassbox software.

Glass Box

Glass Box is a system being developed by the government to provide an instrumented infrastructure for research activity in the intelligence community. The focus is on understanding workstation activity of an intelligence analyst and the

[4]For example, TechSmith's Morae product (www.techsmith.com/morae) records screen video with a list of synchronized system events and Etnio (www.ethnio.com) designed by one of our former students supports remote recording of workstation activity.

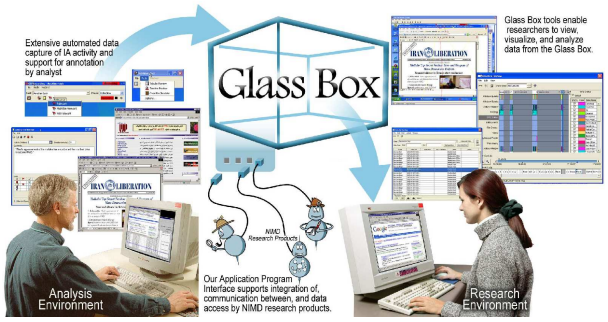

Figure 12: *Glass Box Analysis and Research Environments.*

general analytic processes that happen online. Glass Box is a component in a larger effort of the Advanced Research and Development Activity (ARDA) to "create a new generation of analytic tools to support human interaction with information." [18] The software provides automated data capture of time-stamped activities such as browser activity, active applications, file activity, text that is copied/pasted, keyboard/mouse activity, and screen captures that provide "over-the-shoulder" video of activity. Also provided are review tools for researchers to view and filter data as well as an API to enable access to a central data repository as well as integration of experimental components. It is starting to be used by a number of research efforts (e.g., a Parc effort employs it to study web browsing activity). We have recently downloaded the software and are in the process of installing it in our lab.

We will integrate Glass Box software into our developing visualization and analysis framework and explore use of the TeraGrid as a storage infrastructure to facilitate access by the SIDGrid community. We will record workstation activity while participants are conducting their normal activities and evaluate automatic annotation of the resulting video data. One of our students, Gaston Cangiano, is collecting workstation activity in a legal office with the goal of automatically identifying meaningful events in participants workstation activity. We will focus first on this data. In addition, we will also collect activity data while participants are engaged in other tasks (e.g., work on editing a paper, analyzing data, or creating a graph or figure for including

in a publication) that arise as part of their normal work activities.

Our main focus will be on segmenting video data (captured from workstation framebuffers) into meaningful units, multiscale visualization of those components on timelines, and identification of landmarks in the activity structure. We will use both video annotations and access to application activity provided by Glass Box facilties as a basis for segmentation and landmark identification. Our plan is to compare participants' segmentation and landmarks with both analysts' and those we create automatically.

A.2.4 Writing Activity

Opportunistic use of software from another project [37] will enable us to collect time-based records of the writing process. Francois Guimbretière of the University of Maryland, PI Hollan, and their collaborators have developed PapierCraft [37], a gesture-based command system to support interactive paper. It is based on Guimbretière's earlier *Paper Augmented Digital Documents (PADD)* work [23]. The motivation of both systems is to explore putting the digital and paper worlds on equal footing, one in which paper and computers are simply different ways to interact with documents during their lifecycle. When paper affordances are needed, a document is retrieved from a database and printed. The printer acts as a normal printer but adds a pen-readable pattern to each document. Using a digital pen, the document can then be marked like a normal paper document. The strokes collected by the pen are combined with the digital version of the document. The resulting augmented document can be edited, shared, archived, or participate in further cycles between the paper and digital worlds.

Digital pens provides access to time-stamped records of every pen stroke. This time information is used in our PapierCraft project to help disambiguate gestures but for the current work it can provide histories of the writing process. The visualization of such histories will force our developing analysis framework to confront a different form of time-based data.

While it is beyond the scope of the work we propose here, we envision a variety of future applications based on access to and analysis of encapsulations of the history of the writing process. This may open up exciting research and application possibilities.

A.2.5 Reestablishing Context of Activity

In recent work we asked participants from our driving experiments to return to the lab to view first-person video records of portions of their driving activity as part of extended interviews. We were struck by the vividness of participants' recall while viewing video of their own activity. Saadi Lahlou, one of our collaborators[5] has reported similar findings. He mentions that one of the most remarkable properties of viewing subcam video is "an exceptional capacity of recollection. Even weeks or months after the recorded episode, subjects are able to remember emotions and intentions."

We were similarly struck by a recent report from the first in-depth study[27] of the SenseCam, a novel sensor-augmented wearable still camera. Investigators here too found startling improvements in recall with first-person video. In this case it was part of a 12-month clinical trial with a patient suffering from memory problems originating from brain injury. Not only did review of images show decided advantage over review of a detailed diary but memory actually *improved* during periodic reviews and was maintained for months afterward. Memory was lost in days with no aid or even with review of a detailed diary.

While, as mentioned earlier, there is much work on video summarization, to our knowledge none of this is based on first-person video such as that we are examining. As part of a speculative addition to our main research we will explore how reviewing segments of video from earlier workstation and writing activity might help reestablish the context of those activities. We will compare conditions in which no aid is presented

[5] Saadi Lahlou directs the Laboratory of Design for Cognition at EDF in Paris and is the originator of the Subcam, a personal video recorder, that we have used in our driver studies. The subcam is a small video camera mounted to glasses and connected to a mini-dv portable battery operated video recorder.

with those in which short video clips are viewed. The video clips will either be created by an experienced videographer (Adriene Jenink, Visual Arts professor at UCSD, has agreed to collaborate with us and also assist in recruiting other videographers) using advanced video editing facilities, by our research team using Diver to create clips, and by a variety of automatic techniques we will be exploring as a basis for automated timeline annotation of the video data.

A.2.6 Management Plan

Our project management plan has two dimensions: four activity domains (flight simulator, workstation, writing, and reestablishing context) by three tool building components. We will focus on flight simulator and workstation activity in year one, adding writing and reestablshing context in the out years. Each of the three components to tool building will be led by one of the PIs: (1) Exploration of the application of computer-vision techniques, (Movellan) (2) Development of multiscale visualizations (Hollan) and, (3) Coordinating analysis across multiple levels (Hutchins). Hollan will be in charge of overall management of the project. Because the PIs are located on the same campus, it is easy to arrange frequent meetings with students (graduate and undergraduates) and other local participants. In the first two months of the project we will have a series of planning meetings and retreats in which we will match the needs of the project components with the abilities and interests of participants. Since the world of software is changing so rapidly, we will also use these meetings to adapt our plans to take best advantage of the available technologies.

Through our experience in a fundamentally interdisciplinary department and laboratory we have learned how to establish and maintain a research culture in which our students appreciate different academic cultures (e.g., engineering and anthropology) and work together in an atmosphere of mutual respect. In the planning meetings we will also set milestones for each project component, and for the integration of the components. The overall milestone structure will include publication each year of at least one paper for each tool building component.

Once work is underway, we forsee periodic meetings at three levels of integration. Within each project component, PIs and students will meet very frequently to do the work. Meetings of the assembled project teams will take place at least monthly. Meetings with our user communities (local in person and extended via teleconference) will take place quarterly, at first to update users on development progress and to identify requirements, later as part of the on-going evaluation of their use of newly created analysis facilities.

Project management will also benefit from outside advice. We are well connected to an international network of behavioral scientists who are exploring the organization of real-world human activity (Stanford:Pea, UCLA:Goodwin, Nijmegen:Levinson, Paris:Lahlou). By keeping the community abreast of our efforts, we expect the community to be able to learn from our successes as well as our failures. By sharing tools and ideas with this community, we expect to be able to better understand the nature of the stumbling blocks that impede progress in behavioral sciences as well as to better understand the role that our evolving multiscale analysis framework can play in removing these stumbling blocks. We are also fortunate to be able to seek advice from RUFAE (Research on User Friendly Augmented Environments), a newly formed international network of research institutions (see `http://www.rufae.net` involved in designing and studying interactive spaces. We will draw on the expertise of this group as a sounding board for advice about privacy, evaluation, and other issues that arise. We expect Saadi Lahlou to be an especially valuable research contact. He directs one of the largest labs in Europe devoted to video-based analysis of group interaction. Also Lahlou and Norbert Streitz, another Rufae member, lead a group making recommendations on privacy to the European research community. Privacy policies and issues are central to our proposed research and we expect them to be an important management issue as well. One of the expected outcomes of our project will be documentation of best practices in using digital video in research.

B References

[1] Dan Bauer, Pierre Fastrez, and Jim Hollan. Compuationally-enriched piles for managing digital photo collections. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*, 2004.

[2] Dan Bauer, Pierre Fastrez, and Jim Hollan. Spatial tools for managing personal information collections. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[3] R. Bauman and J. Sherzer. *Explorations in the Ethnography of Speaking*. Cambridge Univesity Press, 1989.

[4] Ben B. Bederson, James. D. Hollan, Ken Perlin, Jon Meyer, David Bacon, and George Furnas. Pad++: A zoomable graphical sketchpad for exploring alternate interface physics. *Journal of Visual Languages and Computing*, 7:3–31, 1996.

[5] Benjamin B. Bederson and James D. Hollan. Pad++: A zooming graphical interface for exploring alternate interface physics. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 17–26, 1994.

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH'99 conference proceedings*, pages 187–194. ACM, 1999.

[7] E. Boer, D. Forster, C. Joyce, P. Fastrez, J.B. Haue, M. Chokshi, E. Garvey, T. Mogilner, and J. Hollan. Bridging ethnography and engineering through the graphical language of petri nets. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[8] Ron Brinkmann. *The Art and Science of Digital Compositing*. Morgan Kaufmann, San Francisco, 1999.

[9] Rodney Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[10] John S. Brown, Allan Collins, and Paul Duguid. Situated cognition and the culture of learning. *Educational Researcher*, 18(1):32–41, 1989.

[11] S. Card, J. D. MacKinlay, and B. Shneiderman, editors. *Readings in Information Visualization*. Morgan Kaufmann, 1999.

[12] S. Chaiklin and J. Lave. *Understanding Practice*. Cambridge University Press, New York, 1996.

[13] William Clancey. *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge University Press, 1997.

[14] Andy Clark. *Being There: Putting Brain, Body and World Together Again*. MIT Press, 1997.

[15] Andy Clark. *Mindware*. Oxford University Press, 2001.

[16] Andy Clark. *Natural-Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*. Oxford University Press, 2003.

[17] Michael Cole. *Cultural Psychology: A Once and Future Discipline*. Belknap Press, 1996.

[18] Paula Crowley, Lucy Nowell, and Jean Scholtz. Glass box: An instrumented infrastructure for supporting human interaction with information. In *Proceedings of the Hawaii International Conference on System Sciences*, 2005.

[19] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. Mylifebits: fulfilling the memex vision. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 235–238, New York, NY, USA, 2002. ACM Press.

[20] Susan Goldin-Meadow. *Hearing Gesture: How our hands help us think*. Belknap: Harvard University Press, 2003.

[21] Charles Goodwin. Professional vision. *American Anthropologist*, 32:1489–1522, 2000.

[22] Charles Goodwin and J. Heritage. Conversation analysis. *Annual Review of Anthropology*, 96:283–307, 1990.

[23] François Guimbretière. Paper augmented digital documents. In *Proceedings of the ACM Symposium on Suer Interface Software Technology*, pages 51–60, 2003.

[24] J. J. Gumperz and D. Hymes. *Directions in Sociolinguistics*. Blackwell, Oxford, 1986.

[25] Jon Helfman and James Hollan. Image representations for accessing and organizing web information. In *Proceedings of the SPIE International Society for Optical Engineering Internet Imaging II Conference*, pages 91–101, 2001.

[26] William C. Hill, James D. Hollan, Dave Wroblewski, and Tim McCandless. Edit wear and read wear. In *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*, Text and Hypertext, pages 3–9, 1992.

[27] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. Sensecam: A retrosepctive memory aid. In *Proceedings of Ubicomp 2006*, pages 177–193, 2006.

[28] J. Hollan, E. Hutchins, and D. Kirsh. Distributed cognition: Toward a new theoretical foundation for human-computer interaction research. *ACM Transactions on Human-Computer Interaction*, pages 174–196, 2000.

[29] I. Hutchby and R. Wooffitt. *Conversation Analysis*. Polity Press, 1998.

[30] E. Hutchins, S. Nomura, and B Holder. The ecology of language practices in worldwide airline flight deck operations. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, July 2006.

[31] Edwin Hutchins. *Cognition in the Wild*. MIT Press, Cambridge, 1995.

[32] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: Systems biology. *Annual Review Genomics – Human Genetics*, 2:343–372, 2001.

[33] V. Jones and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[34] George Lakoff and Mark Johnson. *Philosophy In the Flesh: The Embodied Mind And Its Challenge To Western Thought*. Basic Books, 1999.

[35] J. Lave. *Cognition in Practice: Mind, Mathematics and Culture in Everyday life*. Cambridge University Press, 1988.

[36] Ying Li, Tong Zhang, and Daniel Tretter. An overview of video abstraction techniques. Technical report, HP Laboratories Palo Alto, 2001.

[37] C. Liao, F. Guimbretière, K. Hinckley, and J. Hollan. Papiercraft: A gesture-based command system for interactive paper. *ACM Transactions on Computer-Human Interaction*, page in press, 2007.

[38] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joel Chenu, Takayuki Kanda, Hiroshi Ishiguro, and Javier R. Movellan. Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. In *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, Massachusetts, in press.

[39] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, in press.

[40] T. K. Marks, J. Hershey, J. C. Roddy, and J. R. Movellan. 3D tracking of morphable objects using conditionally gaussian nonlinear filters. Submitted.

[41] Tim K. Marks, John Hershey, J. Cooper Roddey, and Javier R. Movellan. 3d tracking of morphable objects using conditionally gaussian nonlinear filters. *Computer Vision and Image Understanding*, Submitted.

[42] J. McCall, A. Achler, M. Trivedi, J. Haue, P. Fastrez, J. Forster, and J. Hollan. A collaborative approach for human-centered driver assistance systems. In *Proceedings of IEEE Conference on Intelligent Transportation Systems*, 2004.

[43] David McNeil. *Gesture and Thought*. Chicago University Press, 2005.

[44] A. W. Murray. Whither genomics? *Genome Biology*, 1(1):31–36, 2000.

[45] Bonnie Nardi. *Context and Consciousness: Activity Theory and Human-Computer Interaction*. MIT Press, 1996.

[46] S. Nomura, E. Hutchins, and B. Holder. The uses of paper in commercial airline flight operations. In *Proceedings of the 20th Annual Computer Supported Cooperative Work Conference*. ACM SIGCHI, November 2006.

[47] Donald A. Norman. *Things that Make Us Smart*. Addison-Wesley, 1993.

[48] Roy Pea, Michael Mills, Joseph Rosen, Ken Dauber, Wolfgang Effeslberg, and Eric Hoffert. The diver^{TM} project: Interactive digtial video repurposing. *IEEE Multimedia*, 11(1):54–61, 2004.

[49] Roy D. Pea. Practices of distributed intelligence and designs for education. In G. Salomon, editor, *Distributed Cognitions*, pages 47–87. Cambridge University Press, 1993.

[50] C. L. Prvigano and P. J. Thibault. *Discussing Coversation Analysis: The Work of Emanuel A. Schegloff*. John Bengamins Publishing, 2003.

[51] B. Rogoff. *The Cultural Nature of Human Development*. Oxford University Press, New York, 2003.

[52] Lucy Suchman. *Plans and Situated Actions*. Cambridge University Pres, 1987.

[53] E. Thelen and L. Smith. *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, 1994.

[54] F. Varela, E. Thompson, and E. Rosch. *The Embodied Mind*. MIT Press, Cambridge, 1991.

[55] Paul Viola and Michael Jones. Robust realtime object detection. *International Journal of Computer Vision*, 2002.

[56] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.