Appears in *Proceedings Graphics Interface '92,* 1992, 157-166.

Designing Video Annotation and Analysis Systems

Beverly L. Harrison
Department of Industrial Engineering and
Dynamic Graphics Project, Computer Systems Research Institute
University of Toronto

Ronald M. Baecker
Department of Computer Science and
Dynamic Graphics Project, Computer Systems Research Institute
University of Toronto

Abstract

Although video has been used for many years to record data, few tools have been developed to help analyze video data. Current multimedia interfaces have severe cognitive and attentional limitations, reflecting technology-centred designs which have not profited from human factors theory and user-centred design. This paper discusses user requirements for video analysis, from which we derive a set of functional specifications. These specifications are useful for evaluating existing systems and for guiding the development of new systems. A number of existing systems are briefly described, as is the VANNA Video ANNotation and Analysis system, which integrates video, non-speech audio, voice, textual and graphical data, and which incorporates emerging technology, user-centred design and human factors theory.

Keywords

Video annotation, video analysis, analysis methodologies, usability testing, multimedia.

Introduction

Video is a vivid and compelling way of presenting information, of highlighting interesting experimental findings, and of illustrating unique concepts. It provides a highly-detailed, permanent record which can be analyzed in many ways to extract a variety of different types of information. These benefits have long been recognized in usability testing, training and education (e.g., Anacona, 1974; Dranov, Moore, and Hickey, 1980; Ramey, 1989; Nielsen, 1990). New applications in video mail (e.g., Buxton and Moran, 1991), interactive multimedia systems (e.g., Ishii, 1990; Mantei, Baecker, Sellen, Buxton, Milligan, and Wellman, 1991), behavioral research and computer supported cooperative work (e.g., Greif, 1988; Baecker, 1992) are driving an increasing demand for better tools for handling and analyzing video.

Consider the following example. In one typical video analysis application experimenters video taped two subjects in a collaborative programming task. The subjects communicated using a video link between their two distant locations. Experimenters were interested in noting when subjects were looking at the computer monitor and when they were looking at the video monitor. They additionally wanted information about whether subjects looked or did not look at the video image of their partner at times when they communicated. Finally, experimenters wanted to know what types of information were communicated and passed between the two subjects using the video link (e.g., diagrams, pointing to parts of the computer screen or manual, normal conversational gestures). The data resulting from such an analysis might include gaze information as it related to conversation, and data about information content.

For video analysis tools to succeed they must support a wide variety of tasks and analysis styles while still easily capturing essential data. We need to gain new insights into the way people work with multimedia systems. Little is known about creating new classes of interface which manipulate information having temporal dependencies. We need to apply proven design methodologies, human-computer interaction and human factors theory.

The objective of this paper is to provide the reader with an understanding of design issues for video analysis systems, using the VANNA system as an illustrative implementation of one such system. The first section of this paper describes user requirements for video analysis, enabling us to derive a set of functional specifications for building video analysis tools. These specifications have been applied to evaluate existing “landmark” tools and notation systems (e.g., Rein, 1990; Losada and Markovitch, 1990; Potel and Sayre, 1976; Roschelle, Pea, and Trigg, 1990) and suggest guidelines for the development of new multimedia tools. We then describe the VANNA system, which reflects these guidelines and illustrates a number of unique interface design approaches. Tests results for the VANNA system are presented and design implications are discussed.

User Requirements

Our intent is to provide a multimedia video analysis tool which is easily customized to address the characteristics of the task, the application and the user's personalized style. To achieve this task analyses were performed for multiple users within single application domains and across many different application domains (e.g., usability testing, CSCW, behavioral studies). We also conducted literature reviews and surveys, examined existing systems used for video analysis, and interviewed users of these systems to determine which functionality the systems had in common, which functions were most frequently and least often used, and what the common complaints were. (See Harrison and Baecker, 1991; Harrison, 1991 for detailed discussions of the systems examined.) A summary of the most important results of this work is presented here.

From the task analysis, we derived two key points related to the *process* of manipulating a video document. Users tend to work with video in one of two ways: *annotation* and detailed *analysis*. Annotation implies "note taking." Here users are attempting to capture data in real-time, in highly personalized and abbreviated ways. The annotation task is characterized by high cognitive and attentional demands. Detailed analysis typically occurs after the real-time annotation and does not have the same real-time constraints. In this case the user may make many passes over a given segment of tape in order to capture verbal transcriptions (protocol analyses), behavioral interactions, gestural or non-verbal information. As part of this detailed analysis, users may also wish to run statistical analysis, or summarize data in tables or graphs.

Based on the user interviews and surveys of existing systems, we derived a set of user requirements which support *both* the annotation and the detailed analysis process. These were grouped into four categories: coding the data, analyzing and interpreting the data, user interface and device control, and displaying the data. The *coding* category represents methods for entering the various forms of annotational and analysis data. Elements in the *analysis and interpretation* category are those which related to manipulating pre-recorded data, in order to form conclusions about the nature of the data. The *user interface and device control* category embodies some general principles for building user interfaces of video annotation and analysis systems. Finally, when *displaying the data*, there are several general requirements to guide presentation formats and capabilities.

Coding the Data

There are two kinds of coding: real-time or on-line coding which occurs during annotation, and off-line coding, which occurs during analysis. On-line or real-time coding may be thought of as a subset of the overall coding process, where the video may be viewed playing forward at normal speed only, with no opportunity for review. A restricted set of functions is used, which reflects the real-time constraints and high attentional demands. These capabilities allow the user to mark events (typically with a single button press, mouse click, or keyboard stroke), and allow entry of very short text comments. The user must be able to:

- mark the occurrence of an event
- mark the start and stop points for intervals.

The system must be able to:

- capture keystrokes
- capture subject's computer screen.

The "off-line" coding process requires a more comprehensive set of functions. This stage is characterized by the high usage of speed control and reviewing capabilities, and permits the user to perform detailed coding operations of data including:

- user comments or general observations
- verbal transcriptions of the conversation
- non-verbal information, e.g., gestures
- personality or mood measures.

Our current findings indicate that user comments are typically fairly concise (estimated at less than 200 characters for text, or 2-3 figures for graphics). Verbal transcriptions of the conversation may be word-for-word transcriptions or might simply record specific spoken keywords. Non-verbal or gestural information may be described in a number of ways, including sketches, special coding schemes or symbols, and may even be embedded in the conversational record. This information is the most difficult to represent and is therefore most often subject to encoding schemes, as demonstrated by some of the current notations (e.g., Heath, 1986, 1990). Personality measures or mood assessments are often also ranked and encoded. Most "mood" coding schemes in current tools are based on the Bales SYMLOG system for studying small group dynamics (Bales and Cohen, 1979). Several mood notation systems currently exist though none integrate video directly into the analysis tool (e.g., Losada and Markovitch, 1990; Rein, 1990).

Analyzing and Interpreting the Data

Once the video has been coded, any number of analyses may be applied to the data. The level of analysis is dependent upon the experiment objectives, hypotheses and experimental design, but some general capabilities are summarized below:

- play "next" event
- play "previous" event
- group events
- play entire group
- play loopback i.e., play the same sequence over many times
- keyword searching for text data
- basic quantitative data – frequencies, averages, durations, variances
- time series or interaction analyses
- data exporting
- merging of data for interjudge reliability.

Interaction patterns play a significant role in many analyses. These patterns can be derived by statistical means, by time series analysis or by approximation through visual inspection of carefully formatted output. This last case provides a more simplistic view of the interactions by summarizing data on adjacent, aligned time lines. This allows users to visually inspect the data for recurring patterns, overlaps and gaps in interaction. Time lines facilitate the observation of *process* information (as opposed to *content* information).

User Interface and Device Control

The user requirements and the attentional demands of video analysis have direct implications for both the user interface and the mechanisms for device control. Critical interface issues include integration of the video images, device controls, and tool functionality, use of both auditory and visual feedback cues, consistency in the interface, and user-definable screen layouts. Technology issues include the degree of a user's control over the video devices and the choice of input devices.

Integration of the "video monitor" with the annotation system on a single screen is crucial because video requires continuous visual attention; important events might be missed in a split second. This continuous monitoring is required since one can neither predict the frequency of event occurrences, nor the modality for the event (i.e., in which channel the event will occur: auditory or visual). Additionally, spatially separated displays (as are prevalent with existing systems, e.g., Losada and Markovitch, 1990; Roschelle, Pea, and Trigg, 1990) prevent simultaneous access to multiple visual sources. Users must direct their visual attention away from the critical data source in order to locate and select functions in the analysis tool. The resulting visual scan time (and effort) between displays is unacceptable for analysis tasks, in particular for real-time annotation. Finally, integration of displays solves problems with work space size limitations vs. equipment size requirements.

The annotation system should have both auditory and visual feedback mechanisms. If the user is analyzing visual data the auditory feedback cues from the system would be used and vice-versa. This minimizes interference between system feedback and the primary task of analyzing the video data. Visual channels are typically differentiated in terms of spatial separation (i.e., different locations in the visual field). Auditory cues are differentiated by pitch, loudness, and tonal characteristics. The auditory cues should be non-speech to avoid confusion with the voice track of the video document.

In order to successfully record detailed events, comments and information, the user requires automatic control of many of the video speeds from within the analysis tool. The minimum speed control requirements are:

- high speed, e.g., fast forward
- regular playing speed
- frame by frame
- paused at any single frame.

Forward and reverse motion options should be applicable to any of these speeds.

The tool must be capable of coding at a variety of temporal "resolutions". This allows events to be coded at a variety of rates, such as:

- every frame
- every second
- every minute
- at random intervals.

If multiple tapes are used for recording, users may need to automatically cue up any or all of the tapes relative to the position of a single tape.

The coding process, and in particular the real-time annotation, has implications for the style of interface and the input devices chosen. Users need to access the various capabilities of the tool with interfaces which have low visual attentional demands. The kinds of mechanisms might include button presses, touch typing, the ability to point directly to the monitor using a touch screen or draw directly using a stylus. It may be desirable for interface mechanisms and graphic annotations to be overlaid directly on top of the video. The interface should avoid secondary monitors and graphics which require fine motor coordination for function selection. Additionally the mechanisms for representing the data should not require complex encoding schemes or cognitive mappings, but rather should favour a direct one-to-one mapping between the concept to be representing and the interface object (and corresponding label) used to capture the item.

Critical in the usability of any tool is the ability to customize the screen. Users must be able to add or delete instances of objects to reflect their current analysis needs. This includes the ability to modify the tool in an ad-hoc manner during an analysis session. Providing a library of functions from which the users can "copy" and "paste" interface objects and subsequently resize and relocate them on the screen is one method of achieving this.

When reviewing previously recorded data, users must be able easily to play back the previous item, the next item, and user-defined groups of items, independent of their actual location in the data file.

Displaying the Data

The presentation of data and results is perhaps one of the most under-developed aspects of current tools. A minimum requirement is the ability to print a copy of all data recorded. This is basically a "dump" of a log file, containing reference time codes or frame locations, comments, transcriptions, diagrams, event markers, interval markers and keystroke logs. (Many tools do not extend their "output" capabilities much beyond this.) The result is a complex listing of data which is usually so dense that interpretation is difficult. The implication of this is the need for a variety of views or summaries of the results. This includes numerical analyses, time line representations and graphical plots. The user should be able to specify whether the analysis results are presented by experimental subject, by topic discussed, by artifact usage or by other criteria.

Most existing tools present results and data in either tabular format or on a text-based time line (either horizontal or vertical). The use of colour and animation contributes greatly to the clarity and effectiveness of presentation. These techniques have been greatly underutilized thus far.

Often users wish to present short video segments which highlight interesting findings or which provide good representations of general trends in the data. In order to achieve this, they need to be able to indicate the starting and stopping points for a number of sequences, and the order in which the sequences are to be played back. These sequences may be existing intervals marked in the coding stage, or they may be new sequences. The order of presentation of sequences should not be dependent upon the video recording medium, i.e., the tape media must support non-sequential playback.

Functional Specifications

From the user requirements a number of functional specifications may be inferred. These have specific implications for tool functionality and for the user interfaces to video analysis systems.

Coding the Data

1. *User-specified indexing of the video tape.*

Users may mark an event or an interval by indicating the starting (and stopping) position, typically by a button press. Still frames may also be used as index markers. These events or intervals can later be retrieved under computer control.

2. *Grouping of events or activities.*

Users can group similar events or activities together in user-specified classes and assign a unique index to each class.

3. *Experimenter observations or comments.*

Users can enter textual or possibly graphical comments, notes and observations. These are linked automatically to the appropriate segments of video.

4. *Verbal transcript analysis and keyword indexing.*

Users can enter conversational transcriptions of the audio track. Statistics may be computed on the keywords.

5. *Individual and group characteristics.*

Users can enter subjective assessment data for various measures of personality and group dynamics such as Bales measures (Bales, 1950).

6. *Non-verbal and gestural information.*

Users can enter data related to the observed gestural patterns, for corresponding video frames or segments. This information may be coded using a variety of notations, including symbolic notations such as Labanotation (Laban, 1956; Hutchinson, 1954).

Analyzing and Interpreting the Data

7. *Keyword searching.*

Users can use keyword searches on any text data, including experimenter comments or verbal transcriptions.

8. *Keystroke and computer screen integration.*

If the subjects are required to use a computer, their keystrokes and/or computer screen is recorded and synchronized with the video.

9. *Access to text editors, statistics packages, graphics packages, plotting packages.*

Users can import/export data to/from other software packages.

10. *Analysis for interaction patterns over time.*

Users can analyze the data by examining patterns in the occurrence of events or activities over time. Significant (frequent) patterns and interactions are highlighted on a time line.

11. *Support for interjudge reliability.*

A means of merging multiples codings should be available to support multiple judges and hence improve reliability of the data and subsequent analysis.

User Interface and Device Control

12. *Digital control access to basic video functions.*

Users can stop, start, fast forward, and rewind video tape(s) and control the playback speed directly from the analysis tool.

13. *Retrieval and playback of previous and next indexed items.*

Based on the currently selected item or the current location on the tape, users can elect to play the next or previous item recorded.

14. *Retrieval and playback of sets of items using automatic indexing.*

Users can request that *all* events or activities belonging to a given class be played in sequence automatically.

15. *Direct manipulation interface.*

Users access the various capabilities of the tool using interfaces which have low visual attentional demands. The kinds of mechanisms might include button presses, touch typing, the ability to point directly to the monitor using a touch screen or draw directly using a stylus. It may be desirable for interface mechanisms and graphic annotations to be overlaid on top of the video. The interface should avoid secondary monitors and graphics which require fine motor coordination for function selection.

16. *Simplified mental models.*

The users should have minimal mappings and coding schemes to represent events, activities and attributes.

17. *Ability to customize annotation screens.*

Users have access to a “library” of functions, from which a subset may be chosen and laid out on a screen to form user-definable interfaces. Users can relocate and resize any object on the screen.

18. *Multi-media or hypermedia analysis record.*

The final analysis record consists of text, audio, and video is integrated into a single multi-media document.

19. *Automatic synchronizing mechanism for multiple tapes.*

If multiple tapes are used for recording, users can automatically cue up any or all of the tapes based on the position of a single tape.

Displaying the Data

20. *Customizable presentation and summarizing capabilities.*

The number and levels of mappings and coding schemes are minimized to facilitate interpretation of data and results. Results should be presented in user-defined categories. Users specify which items to include or exclude in each summary view. Several standard views are provided. Textual and graphical formats are both available.

21. *Time line display of events.*

Users can view the occurrence of events and duration of activities on time lines. Users can specify the number of time lines and the basis on which they are defined (e.g., per subject, per task, per medium)

22. *Animated and colour displays*.

Animation and movement patterns can be used to illustrate dynamics and capture the temporal dimension of behaviour. Colour can be used to distinguish and highlight variables or interesting results.

23. *Presentation of video segments*.

Users can mark video segments which illustrate relevant or interesting examples and produce an "edit list". This edit list can be easily played back in any sequence.

State of the Art

At the time of this research, there were several interesting video analysis systems in existence, each providing a unique contribution to the field. The tools described may be divided into two categories: notation systems and video analysis tools. Notation systems are methods of representing information extracted from video tape, though they may not be directly linked to or control the video itself. Video analysis tools control the video and integrate functionality such as described in the previous section.

Notation Systems

The Heath Notation System is one of the few encoding schemes which directly integrates detailed information about non-verbal communication and gestures with the corresponding verbal transcripts (Heath, 1986; Heath, 1990). Typed punctuation symbols (e.g., ----, ... , ,,,,,, [,]), represent non-verbal events and activities. Information about intonation, speaking volume and speech characteristics is embedded directly into the transcript, while gestures, gaze and other non-speech information is represented above each line of verbal transcript. This analysis method facilitates observation of the interrelations between non-verbal communication patterns and verbal conversation, although the encoding scheme is complicated and makes accurate keyword searches difficult.

The Mood Meter system is a graphical notation system which is based on the Bales SYMLOG dimensions (Bales and Cohen, 1979) and which describes human interaction and mood over time (Rein, 1990; Olson and Storrosten, 1990). The participants' "mood" ratings are aggregated into a single group score, which is represented diagrammatically by concentric circles or stars of varying colour and density. The idea is to represent divergent groups by dispersed images and convergent groups by concentrated images. One drawback of this tool is the reliance on cognitive mapping schemes for encoding participation and group mood. The mood data require interpretation to convert them to the Bales dimensions, followed by translation to descriptive terms. Additionally, the mood diagrams reflect the *aggregate* group mood, making interpretation on an individual level difficult.

Analysis Tools

GALATEA is "an interactive animated graphics system for analyzing dynamic phenomena [notably biological cell movement] recorded on a movie film" (Potel and Sayre, 1976; Potel, Sayre and MacKay, 1980). In this system computer graphics or animated images are superimposed *directly* on the film. Users have the ability to "write directly" on the film with a digitizing stylus over the video screen image, giving Galatea a unique and truly direct manipulation interface. This input mechanism allows free hand drawing, data point entry, or handwritten notes. There is also a substantial easy-to-use button interface to the video controls which resides on the same monitor as the video image, though is not visible simultaneously.

The GroupAnalyzer is one of the most sophisticated tools for representing group dynamics (mood) over time (Losada and Markovitch, 1990; Losada, Sanchez, and Noble, 1990). The presentation capabilities of this tool are exceptionally good, taking advantage of colour, animation and time series analysis. The analysis component allows users to display both static and dynamic (animated) displays of the results in "field diagrams." Users may display an animation demonstrating how the group dynamics evolve and change over time (the dominance circles for each participant expand and contract). The field diagram may be used to reference the actual video tape. Entering the data requires training, however, since the coding forms are complex and make extensive use of cognitive mappings (per the Bales dimensions). Much of the coding is done in real time by trained experimenters while they are observing subjects.

VideoNoter is a tool which allows users to create annotations and verbal transcriptions which automatically index either a video tape or disc (Roschelle, Pea, and Trigg, 1990; Trigg, 1989). Annotations may be either textual or graphical, and can be composed and subsequently imported from external packages. VideoNoter's particular strength is the interface to the video control functions. It is also one of the few analysis tools which allows users to easily customize their own annotation screen. Users may define their own button oriented coding template for marking events of interest and may reorder the columns by select-and-drag operations. The function of each column is user-definable. Automatic control of video functions are accessible implicitly through the worksheet and scroll bars, or explicitly through a menu bar. User have had to rely heavily on textual entry of data from a keyboard directly into columns in the data file. This time-consuming data coding has restricted the use of this system.

EVA is an interactive video annotation tool which allows both "on-line real-time" coding while the experiment is running, and "off-line" detailed coding with the stored video record after the experiment is competed. (MacKay, 1989; MacKay and Davenport, 1989). It allows experimenters to enter notes, verbal transcriptions, keystroke logging for the subjects, and symbolic categories for organizing the recorded data. One interesting capability of this tool allows the text transcriptions to appear as "subtitles", synchronized with the video. One subtitle appears for each participant. Another facility, not seen in other tools, is the ability to automatically log keystrokes from the subjects and synchronize them with the video. These can be presented in a manner similar to the "transcription subtitles".

U-Test is a tool developed expressly for usability testing and is fairly representative of many usability testing systems (Kennedy, 1989). The emphasis is on "real-time on-line" coding. The tool is pre-programmed, by the experimenter, with a list of tasks that the subjects are to perform. For each task, a detailed list of steps and a set of anticipated possible errors which the subject might make are given. These "error buttons" may be considered specific cases of experimenter-created event indices used to automatically index to specific points on the video tapes. A "timer on" and "timer off" function allows experimenters to set start and stop points for intervals of interest. Experimenters may also enter text comments and observations. These are linked to the video tapes and may be used as indices to specific points on the tapes. The U-Test tool also provides the experimenter with reminders about when they must perform a certain action with regards to the experiment itself.

One of the few tools designed specifically for inexperienced users is Xerox EuroPARC's "virtual VCR" (Buxton and Moran, 1990). A graphical image of a VCR control panel is presented on the user's computer screen. This panel contains all of the standard control functions. In addition to the control functions, the users may "mark" the tape with indices and associated comments (in much the same manner as proposed earlier in this paper). Comments are restricted to short one line titles or notes. These "tags" have a designated start and stop point and a GOTO function for playback.

These systems each address different problems in video analysis. They, and the VANNA system described below, are summarized in Table 1 (shown at the end of this paper) using the functional specifications described earlier.

The VANNA System

The VANNA (Video ANNotation and Analysis) system integrates, on a Macintosh, various multimedia elements into a single video document. User interfaces for the system were created using brainstorming sessions, iterative design, and rapid prototyping, resulting in many versions of the system over a short period of time (approximately 8 versions in 4 months). We used direct manipulation interfaces to support a number of important features described later in this paper. Additionally, we designed the system to support a variety of input devices including a touch screen, digital stylus, mouse, and keyboard. The system supports both real-time annotation and the detailed analysis of video data.

Coding the Data

Users define their own index markers by duplicating index buttons and assigning each a unique name. A single button press immediately creates an index label and links it to the corresponding location in the video document. These indices are used to capture the occurrence of important events in the video and can also be labelled to reflect rankings of behavioral data such as mood and mood changes. Similarly, to capture events having durations, a special index type called an interval is used, indicated using a start/stop button or switch. Users may define any number of indices or intervals. A typical coding screen is shown as Figure 1.

Textual comments may be entered either alone or in conjunction with an index or interval. The comment window is variable length, scrollable and editable. Verbal transcriptions may be thought of as a special case of commenting and are therefore entered in a similar manner. Brief comments of less than 20 characters are typically used for real-time annotation, while lengthy paragraphs and verbal protocols are entered in the more detailed analysis stages. Currently comments may be explicitly linked to any event or interval using a special "link" button.

Analyzing and Interpreting the Data

All annotations are recorded in a log file, with each item type recorded in a different column (e.g., time, indices and intervals, comments). A typical VANNA log file is shown as Figure 2. The log file may be viewed, sorted by any column, edited, searched and played back. Keyword searching and sorting are provided for indices, intervals, comments and transcriptions. Items may be played back by selecting (and hence highlighting) the desired item and pressing the "play" button. Simple built-in data analysis routines calculate the frequency of occurrence for each index label and interval, the average and cumulative durations for each interval, and the variability in duration for each interval. For more detailed statistical analysis or graphical plotting, data may be exported to an external package.

User Interface and Device Control

The VANNA system overview is shown as Figure 3 (at the end of this paper). VANNA simplifies technology access through software interfaces which send and receive video device commands through the Macintosh serial ports. Regardless of video device (e.g., VHS VCR, 8mm VCR, camcorder, video disc) the user sees the same iconic video button controls. Icons are based on the standard video controls found on the devices themselves.

The video image, which may be generated by a video card or by a software solution such as Apple's QuickTime®, appears as a window in the computer monitor. Users may magnify the any portion of the image (zoom in), may shift views of the image (pan), or may pull back from the image (zoom out). This results in a complete integration of video, audio and computer tools, solving many of the problems outlined earlier. The video device(s) may reside in a different room and act as a server(s) to many users. In fact, multiple users may alternately control the same device when working collaboratively. This integration also minimizes work space size requirements (i.e., only the computer workstation is required).

Once created, selected items or groups of items can be played back automatically. (Items may be intervals, indices, comments, or graphics.) The annotation system finds the appropriate location in the video document and begins the playback sequence. This provides users with the ability to easily create “edit lists” of video segments for presentation, based on a number of user defined criteria.

The interface for entering annotations has been designed to reduce both perceptual and cognitive load. VANNA provides users with several default templates or screen layouts and a dictionary of functions. Users may add, delete, resize or relocate any object on a template, including the video window, by directly manipulating the objects themselves (e.g., cut, copy, paste, drag). Only the functions deemed necessary by the user are presented. This creates a completely customizable system.

Multiple input devices are supported simultaneously (touch screen, stylus, mouse, keyboard, video shuttle speed dial). The users may rapidly switch between input devices as appropriate. A shuttle speed dial is available for controlling the video direction and speed. This allows users to take advantage of two-handed input techniques and parallel manipulation strategies (Buxton and Myers, 1986) by controlling the video speed with one hand and pressing buttons with the other. This is particularly useful for detailed analyses when reviewing one segment of video many times at varying speeds to capture information. These devices, combined with the user definable screen layout ability, support both right and left handed subjects equally well.

Any button press provides the user with both auditory and visual feedback. Buttons temporarily inverse highlight and a brief tone sounds (a “clicking” sound like a mechanical button). Different pitches distinguish indices from intervals (in addition to a different visual appearance). The graphics of the interval button changes to differentiate between open and closed intervals. Error tones are louder, with different classes of errors being distinguished by both pitch and tone. Only critical errors display messages, minimizing interference with the primary annotation task.

Displaying the Data

Currently, the system produces reports which display data items in columns. Users define the number of columns and the column content. Columns typically contain the time, index label or interval label, user comments, and verbal transcriptions. The entire contents of the data file may be printed or users may elect to filter the data and print out a pre-determined subset. Additionally, users may view or print out reports using the simple statistics and frequency counters built into the system. Built-in statistics currently include frequency of occurrence for each index and each interval, and the cumulative duration, average duration and variability in duration for each intervals.

User Testing

The VANNA system is undergoing extensive usability testing with a variety of real tasks and applications. These include user interface testing for experiments with pie menus, behavioral studies of writing strategies in joint authoring, studies of gaze patterns for video usage in collaborative programming, usability testing scenarios for several complex devices, studies of human error in kinesiology, and naturally the analysis of video analysis sessions using the VANNA system. A number of interesting results have been observed thus far.

Users rapidly adopted personalized analysis styles, with some favouring real-time annotation and frequent “loopbacks” and others favouring analysis in slow motion with few loopbacks. The button style index markers worked very well and the performance was good for both real-time and non-real-time analysis. In both the real-time and the non-real-time analysis processes users tended to enter comments which were less than 20 characters, though the comments input field was variable length to allow for much longer text items.

Users entered data in sequences of “grouped” button presses. These groupings reflected the data characteristics; some types of events frequently occur in rapid succession. Users physically layed-out the coding screens to reflect these data characteristics by clustering related buttons together. As the data characteristics changed over time, users dynamically altered the screen layout to reflect these changes.

Most users adopted an off-line data coding strategy of entering about 10 items which they then reviewed in the log file for correctness. They would immediately make changes if necessary and then return to the coding process and enter another 10 items. This process resulted in many users requesting a automatically scrolling window view of the most recently entered data as part of the coding screen. By merging a brief view with the coding screen, users felt that their revision and coding process would be simplified. This additionally provides users with feedback about what has been recorded in the data log file.

Comments could be used as data items in themselves or could be descriptors for index markers by explicitly “linking” them. This latter case required the use of a “link” button which has proved to be problematic. Users tended to forget to press this link button when they wished to associate the comment with a specific index marker. Changing the size, location, and label of the link button did not correct the difficulty. A better mechanism for achieving this functionality is needed.

Keyword searching was used extensively in the playback process especially on experimenter comments. The ability to sort by index names and then playback many items from same group was also found to be very useful and was used extensively.

The touch screen was not used extensively but this can be primarily attributed to the angle of the screen, which was found to be too tiring for long sessions. (Most sessions lasted at least 1 hour with an average time of about 2 hours). For future use the touchscreen needs to be mounted in a recessed surface at about a 35 degree angle.

Although video zoom was provided the performance was too slow. Users wanted to rapidly magnify and later demagnify portions of the video image. Better video technologies now make this possible.

Summary and Conclusions

The VANNA system was designed by applying proven methodologies in HCI and human factors theories. It illustrates one method of achieving a cost-effective and useful desktop video annotation and analysis system. Preliminary results in user testing indicate that the VANNA system is suitable for a number of applications for a number of users. Insights from user testing have encouraged us to contemplate a number of extensions.

For example, we have recently implemented a portable version of the annotation subsystem which runs on a laptop computer, the PowerBook. Portable Vanna can be taken into the field and allows real-time annotation to occur simultaneously with data capture.

Graphical overlay capabilities are under implementation. This will allow users to draw sketches using a stylus directly over the video image. We believe that this will prove a useful mechanism for capturing non-verbal and gestural data in behavioral analyses.

VANNA will also be linked to an automatic audio tracking system (Sellen, 1992). This system separates, logs, and graphically plots over time audio contributions from up to four meeting participants. This facilitates the analysis of speech patterns such as pauses, interruptions, and simultaneous speech.

Finally, we are currently implementing graphical time line displays and are investigating the use of color displays and animation for more vivid presentations of results.

Acknowledgements

We appreciate the support to our laboratory and to the first author from the Natural Sciences and Research Council of Canada, the Information Technology Research Centre of Excellence of Ontario, the Institute for Robotics and Intelligent Systems of Canada, Bell-Northern Research, Apple Computer, and Xerox Corporation.

References

Anacona, B. (Ed.) (1974). *The Video Handbook.* New York: Media Horizons.

Baecker, R. M. (Ed.) (1992). *Readings in Groupware and Computer Supported Cooperative Work.* Morgan Kaufmann Publishers.

Bales, R.F., (1950). *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley.

Bales, R.F. and Cohen, S.P. (1979). *SYMLOG: A System for the Multiple Level Observation of Groups.* Free Press.

Buxton, W. and Moran, T. (1990). EuroPARC's Integrated Interactive Intermedia Facility (IIIF): Early Experiences. In *Multi-user Interfaces and Applications*. S.Gibbs and A.A. Verrijn-Stuart (Eds). North-Holland. 11-34.

Buxton, W. and Myers, B. (1986). A Study in Two-Handed Input. In *Proceedings of ACM SIGCHI '86.* 321-326.

Dranov, P., Moore, L. and Hickey, A. (1980). *Video in the 80's: Emerging Uses for Television in Business, Education, Medicine and Government.* White Plains, NY: Knowledge Industry Publications.

Greif, I. (Ed.) (1988). *Computer Supported Cooperative Work: A Book of Readings.* Morgan Kaufmann Publishers.

Harrison. B. L. (1991). The Annotation and Analysis of Video Documents. M.A.Sc. Thesis, Dept. of Industrial Engineering, University of Toronto. April 1991.

Harrison, B.L. and Baecker, R.M. (1991). Video Analysis in Collaborative Work. Working paper. Dynamic Graphics Project, Computer Systems Research Institute, University of Toronto.

Heath, C. (1986). *Body Movement and Speech in Medical Interaction*. Cambridge, England: Cambridge University Press.

Heath, C. (1990). Virtual Looking. Rank Xerox EuroPARC, working paper.

Hutchinson, A. (1954). *Labanotation*. New York: New Directions Publishers.

Ishii, H. (1990). TeamWorkStation: Towards a Seamless Shared Workspace. *Proceedings of the ACM CSCW'90 Conference on Computer-Supported Cooperative Work.*

Kennedy, S. (1989). Using Video in the BNR Usability Lab. *SIGCHI Bulletin* 21(2), October 1989, 92-95.

Laban, R. (1956). *Principles of Dance and Movement Notation..* London: Macdonald and Evans.

Losada, M. and Markovitch, S., (1990). GroupAnalyzer: A System for Dynamic Analysis of Group Interaction, *Proceedings of the 23rd Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, 101-110.

Losada, M., Sanchez, P. and Noble, E.E., (1990). Collaborative Technology and Group Process Feedback: Their Impact on Interactive Sequences in Meetings, *Proceedings of the ACM CSCW'90 Conference on Computer-Supported Cooperative Work..*

MacKay, W.E. (1989). EVA: An Experimental Video Annotator for Symbolic Analysis of Video Data, *SIGCHI Bulletin* 21(2), October 1989, 68-71.

MacKay, W.E. and Davenport, G. (1989). Virtual Video Editing in Interactive Multimedia Applications. *Communications of the ACM*, 32(7), 802-810.

Mantei, M., Baecker, R. M., Sellen, A. J., Buxton, W., Milligan, T., and Wellman, B. (1991). Experiences in the Use of a Media Space. *Proceedings of ACM SIGCHI '91.* 203-208.

Nielsen, J. (1990). Big Paybacks from "Discount" Usability Engineering. *IEEE Software*, 7(3), May 1990. 107-108.

Olson, J.S. and Storrosten, M. (1990). Finding the Golden Thread: Representations for the Analysis of Videotaped Group Work. University of Michigan, working paper.

Potel, M.J. and Sayre, R.E., (1976). Interacting with the Galatea Film Analysis System. ACM Computer Graphics 10(2), July 1976, 52-59.

Potel, M.J., Sayre, R.E. and MacKay, S.A., (1980). Graphics Input Tools for Interactive Motion Analysis. *Computer Graphics and Image Processing* 14, 1-23.

Ramey, J. (1989). A Selected Bibliography: A Beginner's Guide to Usability Testing, *IEEE Transactions on Professional Communication*, 32(4), December 1989.

Rein, G. (May, 1990). A Group Mood Meter. *MCC Technical Report*.

Roschelle, J., Pea, R. and Trigg, R. (1990). VideoNoter: A Tool for Exploratory Video Analysis. *IRL Technical Report* No. IRL90-0021, March 1990.

Sellen, A. J. (1992). Speech Patterns in Video-Mediated Conversations. *Proceedings of ACM SIGCHI '92*, to appear.

Trigg, R. (1989). Computer Support for Transcribing Recorded Activity. *SIGCHI Bulletin* 21(2), October 1989.

Functional specifications	Galatea	GroupAnal.	VideoNoter	EVA	U-Test	Virt. VCR	VANNA
1. User-specified indices	+	+	+	+	+	+	++
2. Grouping items					+		++
3. Experimenter comments	+	++	++		+	+	+
4. Verbal transcription		+	+	+	+		*
5. Characteristics measures		++				+	
6. Non-verbal information	+		+	+			*
7. Keyword searching		+	+	++	++	+	++
8. Keystroke integration			+	++	++		+
9. Import/export data	+	+	++	++			+
10. Interaction patterns		++					
11. Interjudge reliability		++	+	++	+		
12. Video controls	++	+	++	+	+	++	++
13. Previous/next item	++	+	++	+	+	+	++
14. Sets of items			?	?	+	+	+
15. Direct manip. interface	+	+	+	+	+	+	++
16. Simple mental models	++	+	++	+	+	++	++
17. Customizable screens	++		++	+	++		++
18. Multi-media document	++	+	+	++	+	+	++
19. Synchronizing tapes				+	+		
20. Customizable presentation	+	++	++	+	+		+
21. Time line displays		++	+	?			*
22. Anim/colour displays	+	++					*
23. Video presentations	+	+	+	++	++	++	++

Table 1. Assessment of Video Analysis Tools

Legend: **++** Superior capability ***** Planned but not yet implemented
+ Basic capability **?** Information not available
Capability not in system (Blank entry)