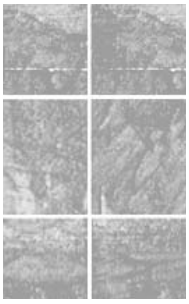
Commonplace Tools for Studying Commonplace Interactions: Practitioners' Notes on Entry-Level Video Analysis

Dan Loehr and Lisa Harper
Visual Communication 2003 2: 225
DOI: 10.1177/1470357203002002006

visual communication



METHODS REVIEW

Commonplace tools for studying commonplace interactions: practitioners' notes on entry-level video analysis

DAN LOEHR AND LISA HARPER
The MITRE Corporation and Georgetown University, Washington, DC

ABSTRACT

The advent of two technologies – inexpensive video recorders and digital video annotation software – is revolutionizing the study of human interaction. With palm-held camcorders, researchers can now easily collect video data in the field, where humans interact most naturally. More important, using free digital video annotation software, one can annotate and analyze video data on a laptop. In this methods review, we discuss our own experience with such tools in the study of gesture and language. We note points of general interest to researchers of other disciplines, and conclude with a discussion of new issues raised by the presence of readily available multimedia data.

KEYWORDS

digital • gesture • human • interaction • multimedia • speech • tool • video

INTRODUCTION

In the middle of the 20th century, the development of inexpensive quality audio tape recorders revolutionized the variety of fields studying human interaction. Researchers could easily and accurately collect audio data in the field, where humans interact most naturally. In the same spirit, the start of the 21st century is witnessing the arrival not only of inexpensive quality video recorders, but also of digital methods for more easily analyzing video. These technologies are having a similar positive effect on the study of nonverbal interaction. Commonplace interactions can now be captured by an inexpensive palm-held video camera, and analyzed with free software on a laptop computer. In this methods review, we discuss our experience as practitioners in using entry-level video hardware and software for linguistic analysis of speech and gesture.

Vol 2(2): 225–233 [1470-3572(200306)2:2; 225–233;032597]

Video technology has been around for over a century. Until recently, however, it has been cumbersome to use for fine-grained analysis of human behavior. In the 1870s, Eadweard Muybridge used multiple cameras with timed shutter releases to capture humans in motion (Muybridge, 1887). The resulting series of photographs could be viewed in rapid succession in Muybridge's 'zoopraxiscope'. In the early 1940s, David Efron (1972[1941]), in his pioneering study of gestures, projected motion-picture films of his subjects frame by frame onto graph paper, upon which he then laboriously traced movement of the hands and arms. In the 1950s, a number of researchers used slow-motion film projectors to analyze interactions, the most famous project arguably being *The Natural History of an Interview* (McQuown, 1971). In the 1960s, William Condon (see, e.g., Condon and Ogston, 1966) refined the use of a hand-operated sound film projector, used in conjunction with a time-aligned oscilloscope, for frame-by-frame 'linguistic–kinesic microanalysis', a method later used extensively by Adam Kendon (see, e.g., 1972, 1980). This method of mechanically advancing frames on a film projector or video cassette recorder (VCR) has been used until very recently. For example, David McNeill's (1992) seminal work on gesture used a VCR with slow-motion and freeze-frame capabilities, and video tapes with frame number and optional oscilloscope traces added.

Apart from the mechanical difficulties of analyzing video in this way, there has been another drawback. This drawback has been shared not only by audio analysis, but also by most scientific disciplines throughout history. The drawback is the fact that researchers have had to transcribe annotations of raw data on a medium separate from the raw data itself. While usually unavoidable, this creates several problems. First, information that is easily perceptible on the raw data must be manually selected and reproduced elsewhere for fruitful analysis. For example, video annotators have typically recorded on paper dozens of phenomena, from movement and location of different body parts to speech phoneme data, for each individual video frame. In addition, apart from the inefficiencies incurred in transcribing phenomena that are readily perceptible in the raw data, there is also selective information loss. Only certain phenomena are siphoned off from the raw data to the analysis stream. If the annotator's transcriptions could co-exist with the raw data, the analyst could quickly and intuitively perceive phenomena that may not have originally been selected for analysis.

These two problems – cumbersome mechanical manipulation of video frames, and separation of video footage from annotations – are being overcome by a single technological advancement: digital video annotation software. Such software allows control of video playback, much like a VCR, while providing time-aligned annotation tracks, much like a musical score. As the analyst plays the video forward or backwards, quickly or slowly, the annotation tracks relevant to the current video frame scroll into view. Thus, the annotations can be viewed and stored alongside the raw data.

In short, the development of inexpensive portable video recorders

makes video data *collection* easier, while the development of digital video annotation software makes video *annotation* and *analysis* more efficient and intuitive. Together, these technologies are facilitating a boom in video-based analysis of humans. In the remainder of this review, we discuss our experiences with these technologies in analysis of speech and gesture.

CAPTURING VIDEO AND AUDIO

We captured our video using consumer-quality digital camcorders, which can be purchased for as little as US$500. The audio can also be captured satisfactorily, at sampling rates sufficient for linguistic analysis, using the camcorder's built-in microphone. Using the built-in microphone allows unencumbered, natural recording of subjects. However, with one microphone it can be difficult to differentiate an individual's speech from overlapping speech of others, or from background noise. Because speech was one of our research interests, we chose therefore to use separate lavaliere microphones, one clipped to the lapel of each subject. Such microphones are inexpensive and can be plugged directly into the audio jack on most camcorders. We had a separate camera and lavaliere microphone for each subject, with the microphone fed into the camera pointing at the microphone's bearer. A camera dedicated to each subject allowed us to get the quality video we needed for gesture analysis, while dedicated microphones allowed us to get the quality audio we needed for linguistic analysis. It can also be helpful to have an additional camera capturing all the subjects in one view.

Our set-up used wired microphones, which required us to ask our subjects to remain seated during the filming. We have also used wireless microphones, which use belt-pack radio-frequency transmitters to transmit the sound. While these permit subjects to move about freely, they are more expensive and require careful set-up to avoid interference from local radio-frequency traffic. Regardless of the type of microphone used, it is helpful to have the sound fed into the camcorder, so that the video and audio are automatically aligned on the videotape. We have also captured sound on separate devices (such as digital MP3 recorders), but this introduces an extremely tricky problem of later aligning the video and audio in digital editing software.

The digital video can be transferred to a computer in a variety of ways. A convenient method is with a FireWire cable, supported by Macintosh and many Windows computers, as well as by many digital camcorders. One can of course also capture video with an analog camcorder, and then digitize it by playing the video from a VCR into a computer video capture board, which can cost around US$100. In either case (digital or analog source video), software is required to control the capture process, but this is inexpensive, ranging from free (supplied with the capture hardware) to under US$100.

Digital video can occupy huge amounts of disk space. For example, a sample 15-minute video, uncompressed, took up 6.7 gigabytes. Therefore, most video files are compressed using a codec (compression–decompression scheme). For example, the Cinepak codec, free from Apple Computer, reduced our file sizes to 4 percent of the original, while still maintaining video quality good enough for gesture research. Another technique to reduce disk usage is to only capture to disk shorter video clips of interest, leaving the rest of the footage on tape. For this, it is helpful to have a VCR capable of playing the camcorder's tape, to facilitate repeated playback during initial selection of clips of interest.

In this section and the next, we have tried to present general principles of digital video capture and editing. For an excellent review of specific details and examples of this process, the reader is referred to the TalkBank Project Website (TalkBank, 2002).

ALIGNING MULTIPLE VIDEOS

We have so far discussed how to capture digital video. Before we turn to annotating it, there is another optional step. If there is video of the same interaction from multiple cameras, it can be useful to have the multiple videos synchronized, so that they will play back in synchrony in the annotation tool. At least one annotation tool, SignStream (Neidle, 2000), supports synchronized playback of multiple videos. Most, however, play back only one video at a time. Therefore, one must use video-editing software to temporally align and 'stitch' the multiple camera views side-by-side into a single video consisting of multiple panes. Examples of video-editing software include Apple FinalCutPro (which we used) and Adobe Premier. Such software, between US$500 and US$1000, can be the most expensive part of a researcher's budget. However, as mentioned, it is only necessary if one wants to stitch multiple camera views together, or carry out other editing of the original footage.

For temporal alignment, such video-editing software allows one to lay out multiple video tracks in parallel on a timeline, to shift them forwards or backwards to align, and then to cut them so that they all start at the same instant. As a landmark in the alignment, one must use some event visible in all camera views. A drawback is that the alignment can only be guaranteed to within half a video frame, as there is no simple way of guaranteeing that the frame shutters on different video cameras open at precisely the same instant. At a typical frame rate of 30 frames per second, frames are 33 msec apart, so the maximum delay between two videos would be 16.5 msec. There do exist specialized hardware/software solutions to ensure that all recording devices (video, audio, and other data streams) are time-stamped and synchronized. However, these are expensive and outside the reach of the 'every-day' researcher to whom we write these notes.

Once temporally aligned, the multiple videos can be stitched together

in the same video-editing software, often by simply 'dragging and dropping' them next to each other. The result will be a single video made up of multiple panes or views of the interaction.

ANNOTATING THE VIDEO

We now turn to the interesting part of video-based research: the annotation and analysis. As mentioned, there has been a recent surge in the development of digital video annotation software. There have been several surveys of such tools, by Bigbee et al. (2001), the Linguistic Data Consortium (2001), The International Standards for Language Engineering (ISLE) project (Dybkjær et al., 2001) and Kipp (2002). We now discuss our experience with a particular tool called Anvil (Annotation of Video and Spoken Language) (Kipp, 2001). We emphasize that all of the tools available are useful. However, we restrict our discussion to Anvil, partly because it is an exemplar of the general class of tools, and partly because Anvil was designed for analysis of gesture and language, which is our particular research interest. Anvil is also free, and runs on most computing platforms.

Figure 1 displays a grayscale image of a sample screenshot of Anvil, which appears in color on the computer. The video window contains VCR-like playback controls, including single-frame movement and variable playback speed. The upper left window gives details of program execution.

Figure 1 Typical screen shot of Anvil, from Kipp (2002:13). Used with the author's permission.

The upper right window gives details of the currently selected track and track element (explained later).

The *annotation board* at the bottom contains a musical-score-like layout. The horizontal dimension is time, marked off by successive video frames. The vertical dimension is a series of horizontal *tracks*, or types of information. Users can define their own tracks, and hence define their own phenomena to annotate. Our phenomena included word transcription, gestural types and phases, and intonational elements. The annotation board also contains a vertical red *playback line* running across all tracks. The playback line is time-aligned with the current video frame. As the video is moved forwards or backwards, the playback line follows suit, and vice versa.

To add an annotation element, one first clicks the start location on the annotation board to add the start frame of the element, which is marked with a green vertical line. Then one advances the video (or drags the playback line to the right) as slowly as needed until the end of the desired element. Another click will then bring up a pop-up menu from which to choose the element type. This element will be inserted with the chosen endpoints. Hierarchically larger elements can be attached to groups of more basic ones.

We mentioned that Anvil is tailored for annotation of gesture and speech. Strictly speaking, however, it does not allow acoustic analysis of speech, specifically because excellent speech analysis tools are already available. To meet this need, though, Anvil can import data from a popular speech analysis package, Praat (Boersma, 2001). Like Anvil, Praat is freely available and runs on most computing platforms. Our procedure, therefore, was to annotate the audio track of our video clips in Praat, and then import those annotations into Anvil, to have them side-by-side with the gestural annotations. Anvil can also import a pitch track from Praat, and can generate a waveform display.

Neither is Anvil designed for *statistical* analysis, focusing instead on annotation. However, it can export time-stamped annotations to a file suitable for analysis in a spreadsheet or statistical package like SPSS (SPSS, 2002). Anvil also has a built-in search capability, to easily find and jump to elements of interest.

Anvil is typical of the general class of video annotation tools. Of the eight tools in Bigbee et al.'s (2001) representative survey, the majority used musical-score layouts time-aligned with a video screen, allowed multiple tracks of information, had some search capability, and had some ability to import and export annotations from and to other software.

In this section, we have assumed that the researcher already has an annotation framework in mind for transcribing human interactions. For example, we have used McNeill's (1992) coding scheme for gesture, and the ToBI (Tone & Break-Index) scheme (Beckman and Elam, 1997) for intonation. For those interested in investigating coding frameworks, an excellent survey of 21 multimodal annotation schemes is Knudsen et al. (2002).

MULTIMEDIA DATA ISSUES

The availability of multimedia data brings up a number of issues.

We have already discussed the problem of storage requirements for digital video. This also raises the question of how to share data. CDs must often be used to transfer video files, and soon DVDs may become widespread enough for file transfer. Incompatible codecs can also be a problem. There are currently many different kinds, and a researcher receiving (and decompressing) videos needs to have the same codec as the researcher sending (and compressing) the video.

When sharing video data, confidentiality of subjects becomes more of an issue. Permission must be obtained to show subjects' faces. Alternatively, faces can be blurred or blocked out in the video with digital editing tools.

We have discussed annotation of video and audio. What about other time-based media? A current example is electronic 'ink', or markings made by subjects on a computer or electronic whiteboard. Another example is gaze direction, automatically captured by gaze-tracking devices. Future researchers interested in all aspects of the environment being recorded may even want to track phenomena such as parts-per-million present of certain scents. A tool called MDB-GSG (Multimedia Database-Gesture-Speech-Gaze) (Quek et al., 2000) allows for plotting of any time-based data stream. MDB-GSD can also replay, in an avatar, gaze and hand motion that was automatically captured by special position-calibrated cameras.

Looking to the future, it is interesting that none of the tools for annotating speech and gesture make *use* of speech and gesture. Data entry tasks can be facilitated by speaking and pointing to a computer screen, as shown by Oviatt et al. (1997) in a map-based data-entry task. Therefore, one could conceive of a tool in which, while a video is playing slowly, the researcher could say 'start headshake ... stop headshake' to annotate a headshake. Alternatively, one could point to the timeline and say 'headshake from here <point> to here <point>'.

CONCLUSION

The advent of two technologies – inexpensive video recorders and digital video annotation software – is revolutionizing the fields studying human interaction. Research tools that were once the domain of a few dedicated researchers are now easily available and easily used. We have discussed our own experience with such tools in the study of gesture and language. We hope that these commonplace tools will make video studies of commonplace interactions somewhat, well, more commonplace.

ACKNOWLEDGEMENT

This work has been supported by The MITRE Corporation, under The MITRE Technology Program.

REFERENCES

Beckman, M. and Elam, G. (1997) 'Guidelines for ToBI Labeling, version 3', URL (consulted Nov. 2002): http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/

Bigbee, A., Loehr, D. and Harper, L. (2001) 'Emerging Requirements for Multi-Modal Annotation and Analysis Tools', Proceedings, *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, September.

Boersma, P. (2001) 'Praat, A System for Doing Phonetics by Computer', *Glot International* 5(9/10): 341–5.

Condon, W. and Ogston, W. (1966) 'Soundfilm Analysis of Normal and Pathological Behavior Patterns', *Journal of Nervous and Mental Disorders* 143: 338–47.

Dybkjær, L., Berman, S., Kipp, M., Olsen, M., Pirrelli, V., Reithinger, M. and Soria, C. (2001) 'Survey of Existing Tools, Standards, and User Needs for Annotation of Natural Interaction and Multimodal Data', ISLE Deliverable D11.1, URL (consulted Nov. 2002): http://www.ilc.pi.cnr.it/EAGLES96/isle/

Efron, D. (1972[1941]) *Gesture, Race, and Culture*. The Hague: Mouton. (Originally published as *Gesture and Environment*. New York: King's Crown Press.)

Kendon, A. (1972) 'Some Relationships between Body Motion and Speech: An Analysis of an Example', in Aron Siegman and Benjamin Pope (eds) *Studies in Dyadic Communication*. New York: Pergamon Press.

Kendon, A. (1980) 'Gesticulation and Speech: Two Aspects of the Process of Utterance', in Mary Ritchie Key (ed.) *The Relationship of Verbal and Nonverbal Communication*, pp. 207–27. The Hague: Mouton.

Kipp, M. (2001) 'Anvil – A Generic Annotation Tool for Multimodal Dialogue', Proceedings, *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, September.

Kipp, M. (2002) *Anvil 3.5 User Manual*, URL (consulted Nov. 2002): http://www.dfki.de/~kipp/anvil/doc/Anvil35.pdf

Knudsen, M., Martin, J-C., Dybkjær, L., Ayuso, M., Bernsen, N., Carletta, J., Heid, Ul, Kita, S., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., Van Elswijk, G. and Wittenburg, P. (2002) 'Survey of Multimodal Annotation Schemes and Best Practice', ISLE Deliverable D9.1, URL (consulted Nov. 2002): http://www.ilc.pi.cnr.it/EAGLES96/isle/

Linguistic Data Consortium (2001) 'Gesture Annotation: Tools and Data', URL (consulted Nov. 2002): http://morph.ldc.upenn.edu/annotation/gesture

McQuown, N. A. (ed.) (1971) *The Natural History of an Interview*. Microfilm Collection of Manuscripts on Cultural Anthropology, 15th series, Joseph Regenstein Library, University of Chicago.

McNeill, D. (1992) *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

Muybridge, E. (1887) *Animal Locomotion: An Electro-Photographic Investigation of Consecutive Phases of Animal Movements.* Philadelphia: J.B. Lippincott.

Neidle, C. (2000) 'SignStream™: A Database Tool for Research on Visual-Gestural Language', American Sign Language Linguistic Research Project, Report Number 10, Boston University, URL (consulted Nov. 2002): http://www.bu.edu/asllrp/reports.html#RPT10

Oviatt, S.L., DeAngeli, A. and Kuhn, K. (1997) 'Integration and Synchronization of Input Modes during Multimodal Human–Computer Interaction', Proceedings, *Conference on Human Factors in Computing Systems (CHI '97)*, Atlanta, March, ACM Press.

Quek, F., Bryll, R., Kirbas, C., Arslan, H. and McNeill, D. (2000) 'A Multimedia Database System for Temporally Situated Perceptual Psycholinguistic Analysis', Proceedings, *3rd International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior)*, Nijmegen, The Netherlands, August.

SPSS (2002) 'SPSS', URL (consulted Nov. 2002): http://www.spss.com/

TalkBank (2002) 'TalkBank Digital Video Guide', URL (consulted Nov. 2002): http://www.talkbank.org/dv/

BIOGRAPHICAL NOTES

DAN LOEHR is an Artificial Intelligence Engineer at The MITRE Corporation, conducting research on multimodal interfaces. He is also a PhD candidate at Georgetown University, writing a dissertation on intonation and gesture.

Address: The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102, USA. [email: loehr@mitre.org]

LISA HARPER is an Artificial Intelligence Engineer at The MITRE Corporation, conducting research on multimodal dialogue systems. She is also a PhD candidate at Georgetown University, writing a dissertation on gesture and semantics.

Address: as Dan Loehr [email: lisah@mitre.org]