Behavior Research Methods, Instruments, & Computers
2001, 33 (3), 311-320

SignStream: A tool for linguistic and computer vision research on visual-gestural language data

CAROL NEIDLE, STAN SCLAROFF, and VASSILIS ATHITSOS
Boston University, Boston, Massachusetts

Research on recognition and generation of signed languages and the gestural component of spoken languages has been held back by the unavailability of large-scale linguistically annotated corpora of the kind that led to significant advances in the area of spoken language. A major obstacle has been the lack of computational tools to assist in efficient analysis and transcription of visual language data. Here we describe SignStream, a computer program that we have designed to facilitate transcription and linguistic analysis of visual language. Machine vision methods to assist linguists in detailed annotation of gestures of the head, face, hands, and body are being developed. We have been using SignStream to analyze data from native signers of American Sign Language (ASL) collected in our new video collection facility, equipped with multiple synchronized digital video cameras. The video data and associated linguistic annotations are being made publicly available in multiple formats.

Research on recognition and generation of signed languages and the gestural component of spoken languages has been hindered by the unavailability of large-scale linguistically annotated corpora of the kind that led to significant advances in the area of spoken language. A major obstacle to the production of such corpora has been that annotation of visual language data by human transcribers is extremely time-consuming, even with the assistance of the best available computational tools. This is, in part, because such data must include complex information about simultaneously occurring movements of the hands, the head, and the upper body, as well as facial expressions, which all play an essential role in the grammars of signed languages.

To facilitate the linguistic annotation and analysis of visual language data, we have developed a Mac OS application called SignStream. The program has been widely distributed and is being used to analyze a variety of signed languages; it is beginning to be used for the study of gesture, as well. Our own linguistic research, however, has focused on American Sign Language (ASL; Neidle, Kegl, MacLaughlin, Bahan, & Lee, 2000),[1] and we have been using SignStream to analyze data from native signers of ASL. Most recently, we have been analyzing data that we collected in our new state-of-the-art digital video collection facility, set up as part of the National Center for Sign Language and Gesture Resources (NCSLGR, a collaborative project involving researchers at Boston University and the University of Pennsylvania), which allows us to capture the signing simultaneously from multiple angles. The video data collected in the NCSLGR and the corresponding annotations are made publicly accessible as part of this project. It is our hope that the corpora being produced and distributed will accelerate linguistic and computational research on signed languages and gesture.

Another significant part of this effort is the development of machine vision methods that can be used to assist linguists in many aspects of the video annotation. Efforts are under way to develop machine vision algorithms that can automatically track and estimate head motion, facial movements and eye movements, as well as hand shape, orientation, and trajectory in collected video sequences.

This paper will now describe, in turn, several components of this project, specifically the SignStream application, the computer vision research in progress, and the data collection facility.

SignStream

SignStream, a multimedia database tool distributed on a nonprofit basis to educators, students, and researchers, provides a single computing environment within which to view, annotate, and analyze digital video and/or audio data. SignStream provides direct on-screen access to video and/or audio files and facilitates detailed and accurate annotation, making SignStream valuable for linguistic research on signed languages and the gestural

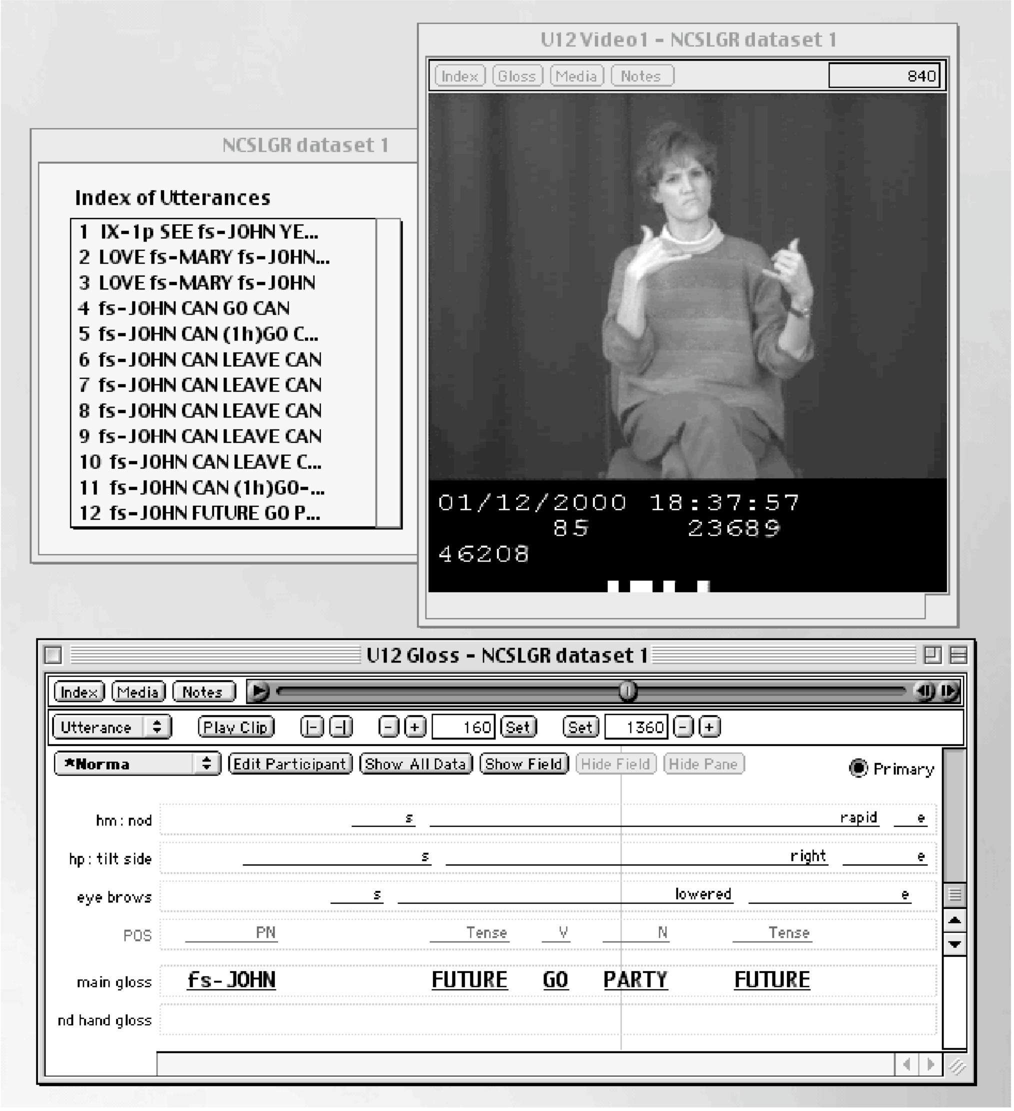
The work reported here has been supported in part by grants to Boston University from the National Science Foundation (BCS-9410562, BCS-9729010, IIS-9528985, IIS-9912573, and EIA-9809340). SignStream software (©1997–2000, Trustees of Dartmouth College, Trustees of Boston University, and Rutgers–The State University of New Jersey) has been developed by Carol Neidle, Dawn MacLaughlin, Benjamin Bahan, Otmar Foelsche, Judy Kegl, and David Greenfield. SignStream programming (through Version 2) has been carried out at Dartmouth College by David Greenfield in the Department of Humanities Resources (Otmar Foelsche, Director). The National Center for Sign Language and Gesture Resources involves collaboration of researchers at Boston University (Carol Neidle and Stan Sclaroff) and the University of Pennsylvania (Dimitris Metaxas, Norm Badler, and Mark Liberman). Correspondence should be addressed to C. Neidle, Department of Modern Foreign Languages, Boston University, 718 Commonwealth Avenue, Boston, MA 02215 (e-mail: carol@bu.edu).

component of spoken languages. Moreover, SignStream may be useful in very different domains involving annotation and analysis of digital video data.[2] Although SignStream currently runs only on Mac OS computers, it is possible to export the coded information in text format. A reimplementation of the SignStream application in Java (with additional functionalities) is now in progress; the next version will also include XML-based import and export capabilities.

Features of SignStream Version 2.0

A SignStream database consists of a collection of utterances, each of which associates a segment of video with a fine-grained multilevel transcription of that video. A database may incorporate utterances pointing to one or more movie files. SignStream allows the user to enter data in a variety of fields, such that the start and end points of each data item are aligned to specific frames in the associated video. A large set of fields and values is provided; however, the user may create new fields or values or edit the existing set. The program aligns information on screen, thus providing a comprehensive visual representation of the temporal relations among linguistic events. As coding changes, the display is adjusted dynamically so as to maintain temporal alignments on screen. A screen shot is shown in Figure 1.

Data may be entered in one of several intuitive ways, including typing text, drawing lines, and selecting values from menus. For example, to enter an item in a gloss-type text field, the user simply types in the gloss, and then, with that item selected, the user can advance the movie to the frame to be assigned as the start point for the sign, and click the "set start" button; the user can then do the same thing to set the end point with the "set



01/12/2000 18:37:57
85 23689
46208

Index of Utterances

1 IX-1p SEE fs-JOHN YE...
2 LOVE fs-MARY fs-JOHN...
3 LOVE fs-MARY fs-JOHN
4 fs-JOHN CAN GO CAN
5 fs-JOHN CAN (1h)GO C...
6 fs-JOHN CAN LEAVE CAN
7 fs-JOHN CAN LEAVE CAN
8 fs-JOHN CAN LEAVE CAN
9 fs-JOHN CAN LEAVE CAN
10 fs-JOHN CAN LEAVE C...
11 fs-JOHN CAN (1h)GO-...
12 fs-JOHN FUTURE GO P...

hm : nod s rapid e
hp : tilt side s right e
eye brows s lowered e
POS PN Tense V N Tense
main gloss fs-JOHN FUTURE GO PARTY FUTURE
nd hand gloss

Figure 1. SignStream video and gloss windows.

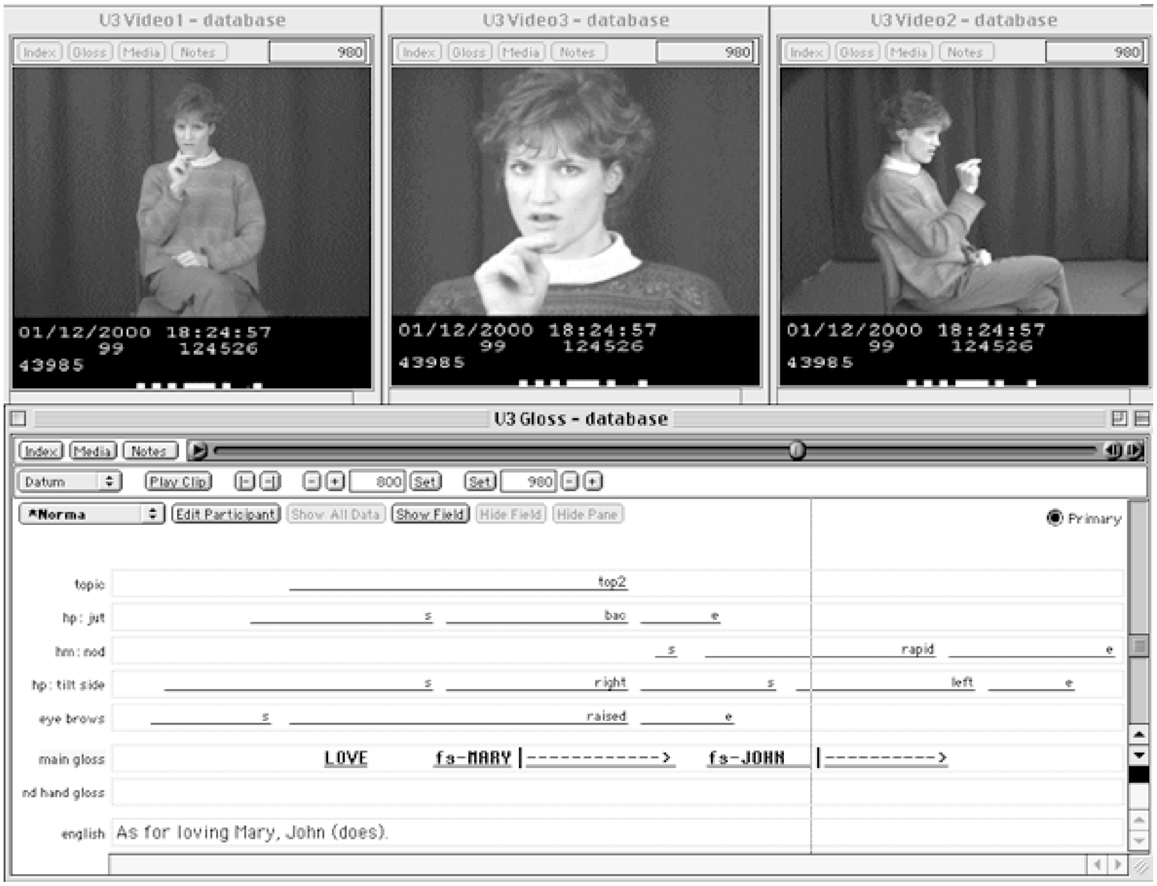
end" button. A nonmanual item can be created by drawing a line in the appropriate field (the movie advances as the user drags the mouse to draw the line), and when the mouse button is released, the user may choose a value from the menu that appears. The screen display (including color and arrangement of fields) can be configured by the user. For any data object that has been defined, including the utterance as a whole, tools are provided to enable the user to go to the start or end frame of that object, to change the start or end point of that object, or to play the video clip associated with that object.

It is possible to display up to four different synchronized video files, in separate windows, for each utterance. A window is also available to display the wave form corresponding to the associated sound file (if there is one). This is shown in Figures 2 and 3. It is possible to step through the media files or to play any segment at the desired speed. The media alignment indicator in both the gloss and audio windows marks the current position (i.e., the frame currently displayed in the video window[s]). The media alignment indicator advances as the video plays, and it is also possible to drag the media alignment indicator in order to shuttle the video forward or backward.

The gloss window may contain multiple participant panes (with all information aligned) to enable coding of multiple participants in a conversation. It is also possible to view distinct utterances (from one or more SignStream databases) on screen simultaneously. Thus, it is easy to compare two different utterances.

SignStream includes an integrated search capability. A SignStream user can formulate complex queries, combining datum specifications using both standard Boolean operators (*and*, *or*, *not*) and specialized temporal operators (*with*, *before*, *after*). For example, one can search for a pointing sign occurring simultaneously with eye gaze. Searches can be conducted in multiple stages, allowing the user to successively refine the search. Queries can be repeated on different search domains, and search files can be saved. In addition, a script facility makes it possible to select a subset of utterances to be viewed in a specific order (with obvious applications for research and teaching). Both searches and scripts can be saved.

In conjunction with the SignStream project, a data repository has been established to enable users to share SignStream-encoded data. The availability of video data (as well as tools for analyzing such data) has the potential to improve linguistic research on signed languages in an important way. A major problem in the field to date has been that data have been reported in the scientific literature primarily through use of an impoverished English-



topic: top2

hp : jut: s bac e

hm : nod: s rapid e

hp : tilt side: s right s left e

eye brows: s raised e

main gloss: LOVE fs-MARY |------------> fs-JOHN |---------->

nd hand gloss:

english: As for loving Mary, John (does).

Figure 2. SignStream display with multiple video windows.

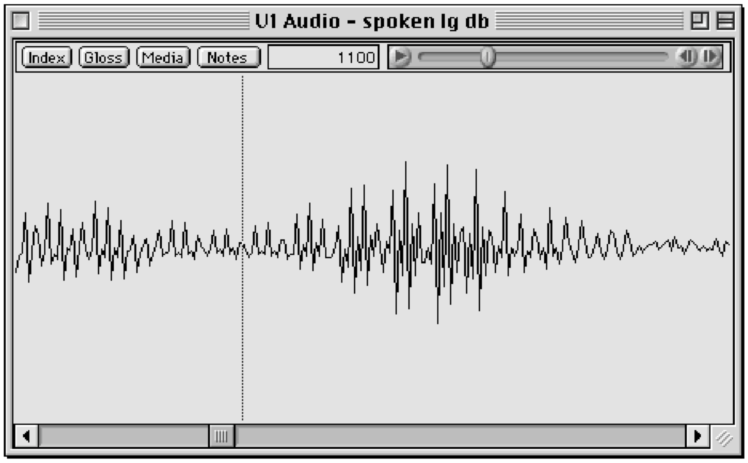


Figure 3. SignStream audio window.

based gloss notation, without access to the primary data. The scientific scrutiny that direct access to data makes possible is essential to progress in the linguistic study of signed languages.

The fact that SignStream is distributed on a nonprofit basis facilitates the sharing of data with other researchers. However, there are other programs available commercially that provide support for analysis of multimedia data, such as Noldus's Observer Video-Pro (described in Noldus, Trienes, Hendriksen, Jansen, & Jansen, 2000, and at http://www.noldus.com), which costs over $5,000 for academic users. The Observer system provides video capture annotation capabilities for the study of human behavior. Through a graphical interface, users can add annotations that are synchronized with the captured video. The system also provides a graphical interface that allows users to view existing annotation files synchronized with video. In contrast to the SignStream system, the Observer system interface does not seem to provide a window that allows viewing of multiple, time-aligned annotation "tracks." Neither does the Observer system provide special-purpose tools to facilitate detailed annotation of gestures. Such tools have proven to be a major time saver in SignStream annotation.[3]

Plans for Further Development

Version 2.0 of SignStream was released in December 2000. That is the version described here. However, SignStream is currently under development. We anticipate that future versions of the program, to be implemented in Java, will incorporate tools for intuitive and efficient entry of fine-grained phonological information. We are now designing a graphical interface that will permit efficient coding of (clusters of) phonological information about such things as hand shape, movement, orientation, location, and so on.[4] Linguistic annotation will be accomplished through intuitive and efficient graphical user interfaces tailored to the kind of information that is being coded.

We also intend to implement the capability to export SignStream databases to XML format. (XML is a flexible and widely recognized standard for data exchange and database connectivity.) Conversely, SignStream will also be able to import data in XML format. This is one way, for example, to import the results of computer vision algorithms, such as those described in the next section. The ability to display graphical representations of numerical data of various kinds in separate windows, with an alignment indicator marking the current frame, will greatly expand the functionality of the program. The kinds of data that could be displayed in this way include movement velocities, repetitive head movements, eyebrow position, and audio characteristics. This capability will make possible several different types of research applications. In some cases, graphical information could be used to aid and/or verify transcription. In other cases, numerical information could be imported and compared with the transcription in order to verify the correctness of the numerical data. Ultimately, importing the output of vision algorithms could result in semi-automation of aspects of the transcription process itself. These advances will result, we hope, in an extremely powerful computing environment within which multiple types of information can be displayed and manipulated, greatly increasing the utility of SignStream for linguistic and computer science research.

For additional information about the SignStream project, see http://www.bu.edu/asllrp/SignStream/ and documents and references found there (especially Neidle, in press, and MacLaughlin, Neidle, & Greenfield, 2000). The developers welcome comments and suggestions about features that might be desirable for different kinds of research.

COMPUTATIONAL RESEARCH IN PROGRESS

As part of this project, we are developing semi-automatic and automatic machine vision algorithms for estimating head motion as well as hand shape, orientation, and trajectory. The goal is to devise computational algorithms that can assist linguists in detailed analysis and annotation of sign language. It is hoped that such tools will enable faster and more accurate annotation, as well as new types of analyses of collected video. In the rest of this section, we provide a brief overview of the effort in machine vision algorithm development to support annotation of hand, head, and facial gesture.

Head and Facial Gesture Annotation Algorithms

The initial focus of the effort has been on machine vision algorithms for estimating 3-D head motion and coding linguisticallysignificant gestures of the head. Although much of prior work in sign language recognition has focused on hand gestures, critical linguistic information is conveyed by positions and movements of the head and upper body, such as eyebrow position, eye gaze, and nods and shakes of the head that occur in parallel with manual signing. It has been estimated that 80% of the grammar of ASL is expressed by nonmanual markings that extend over precise phrasal domains.[5] Thus, no system designed for sign language recognition (or annotation) can afford to overlook this critical component of signed languages.

Therefore, one of our first goals is to develop automatic face detection and head tracking algorithms that estimate the head gesture at each video frame and detect certain linguistically significant periodic gestures in video. Algorithms for determining the 3-D head position and orientation are fundamental in the development of head gesture annotation modules. Furthermore, modules for facial expression analysis, eye gaze and eyebrow tracking, and mouth shape estimation require a stabilized image of the face obtained from a 3-D head tracker that factors out changes of orientation and position of the head.

Several techniques have been proposed for 3-D head motion and face tracking. Some of these focus on 2-D tracking (Crowley & Bérard, 1997; Hager & Belhumeur, 1998; Oliver, Pentland, & Bérard, 1997; Yacoob & Davis, 1996; Yuille, Hallinan, & Cohen, 1992), while others focus on 3-D tracking or stabilization. Some methods for recovering 3-D head parameters are based on tracking of salient points, features, or 2-D image patches (Azarbayejani, Starner, Horowitz, & Pentland, 1993; Jebara & Pentland, 1997). Others use optic flow to constrain the motion of a rigid or nonrigid 3-D surface model (Basu, Essa, & Pentland, 1996; DeCarlo & Metaxas, 1996; Li, Rovainen, & Forcheimer, 1993). More complex models for the face that include both skin and muscle dynamics for facial motion were used in Essa and Pentland (1997) and Terzopoulos and Waters (1993). Global head motion can also be tracked using a plane under perspective projection (Black & Yacoob, 1995). Most of these techniques are not able to track the face in the presence of large rotations or changes in lighting conditions, and some require accurate initial fit of the model to the data; they are, therefore, unsuitable for our application.

Head tracking can, however, be formulated in terms of a 3-D computer graphics model of the human head and face, which circumvents these problems. A complete description of the algorithm appears in La Cascia, Sclaroff, and Athitsos (2000). The method enables fast and stable on-line tracking of extended sequences, despite noise and large variations in illumination. The tracking is made more robust and less sensitive to changes in lighting

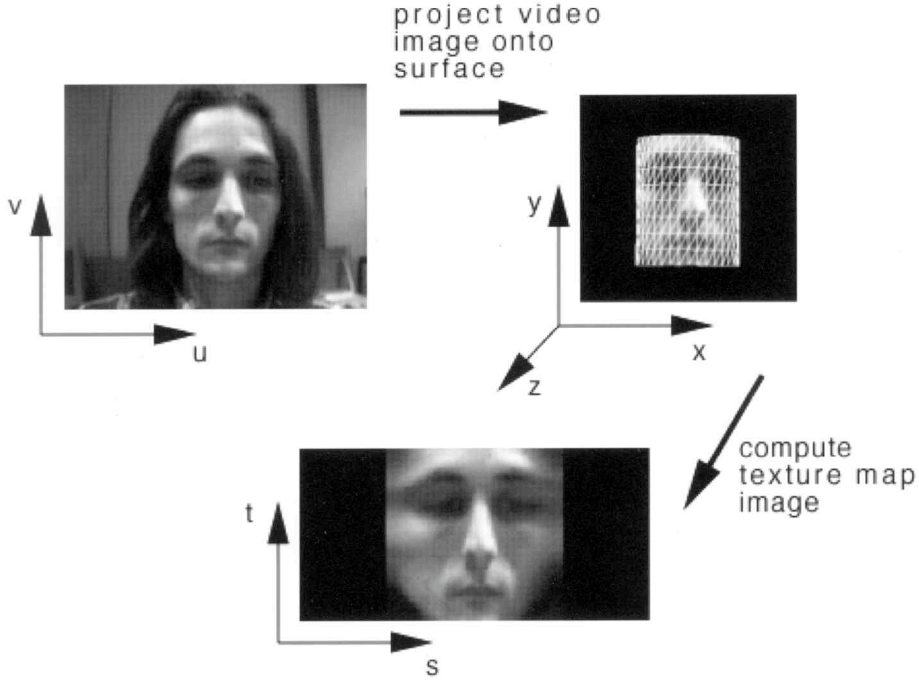


Figure 4. Mapping input video image onto the 3-D cylindrical head model.

through direct modeling of illumination in the computer graphics model. A brief overview of the basic approach, as it relates to the annotation problem, is provided here.

The model we use is a rigid cylinder that is parameterized by its 3-D position and orientation. In order to start the head tracker, the model must be fit to the initial frame. This initialization is accomplished automatically using a 2-D face detector (Rowley, Baluga, & Kanade, 1998) (assuming the subject is facing the camera). After automatic initialization, the subject's face is captured as an image that is applied to the computer graphics model. An example mapping of the input frame onto the computer graphics model is shown in Figure 4.

Once the model is initialized to the first frame in the video, tracking of the head motion over the entire sequence is computed. Tracking is based on changing the location and orientation of the computer graphics model so that the difference between the model and the input video frame is minimized. To account for illumination variations, we adjust the computer graphics model's illumination parameters to account for lighting variations that occur during a video sequence.

The system has been extensively tested (La Cascia et al., 2000). During real-time operation, in many cases, the cylindrical tracker can track the video stream indefinitely—even in the presence of significant motion and out of plane rotations. However, to better test the sensitivity of the tracker and to better analyze its limits, we collected a set of over 70 challenging sequences in which ground truth data were simultaneously collected using a

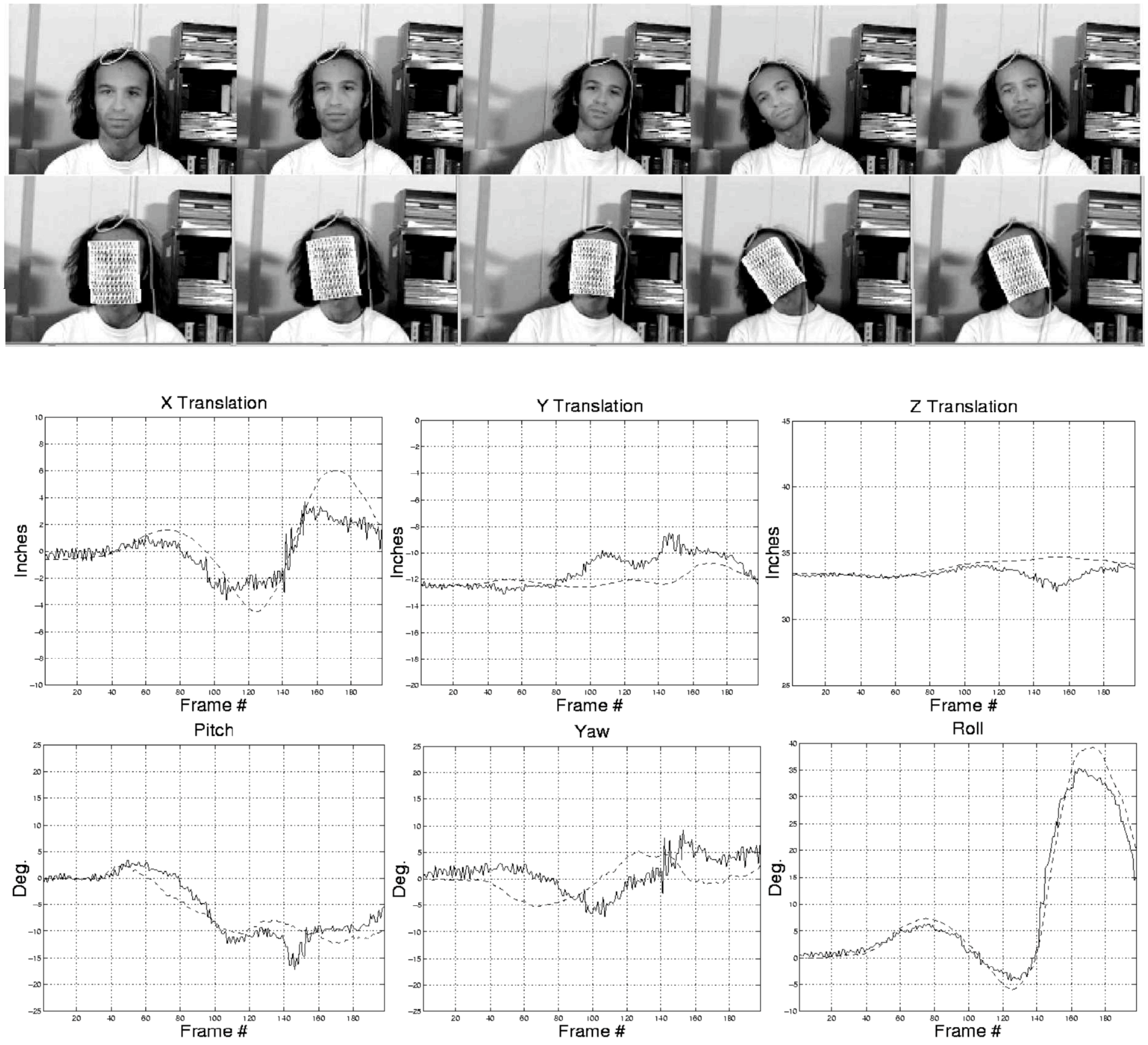


Figure 5. Example head tracking sequence. In all of the graphs, the dashed curve depicts the estimate gained via the visual tracker and the solid curve depicts the ground truth obtained via a magnetic tracker mounted on the subject's head. The first row of graphs shows the *x*, *y*, and *z* translations respectively, where translation is measured in inches. The second row of graphs shows estimates for rotation around the *x*-, *y*-, and *z*-axes, respectively, as measured in degrees.

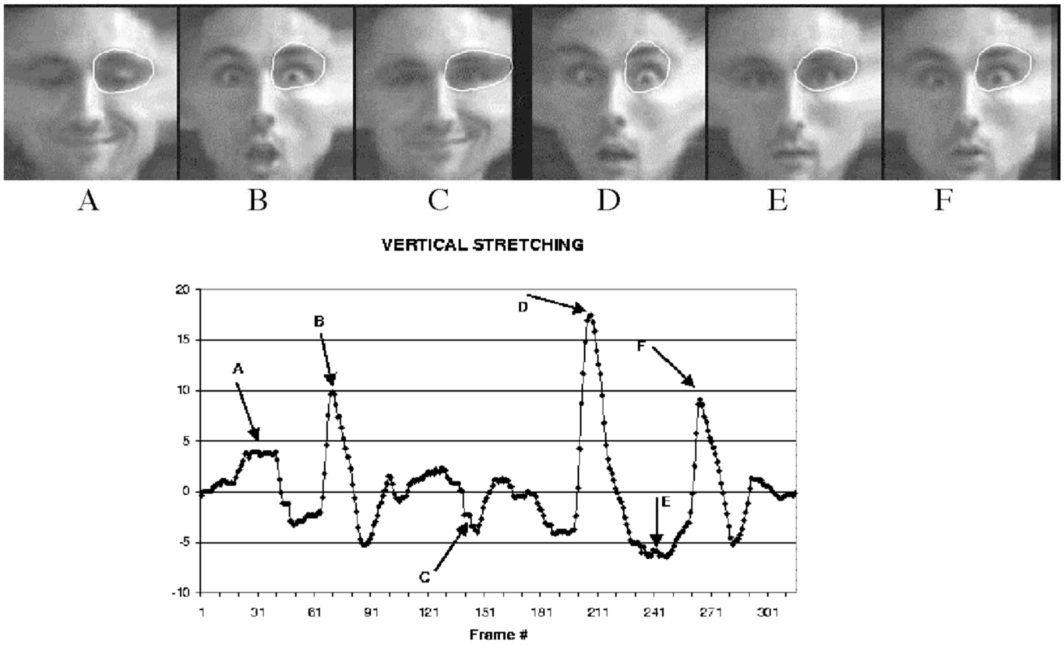


Figure 6. Example of nonrigid tracking in the stabilized dynamic texture map to detect eyebrow raises. The three peaks in the plot correspond to the three eyebrow raises occurring in the sequence.

magnetic tracker. Figure 5 shows an example sequence and the recovered head orientation and translation. The experiments showed that the system is very robust with respect to errors in the initial estimate of the position and scale of the face. The precision and stability of the tracker remain approximately constant for a range of initialization errors of up to 20% of the face size.

The model parameters obtained during tracking determine the projection of input video onto the surface of the object. Taken as a sequence, the projected video images comprise a stabilized view of the face that is independent of the current orientation, position, and scale of the surface model. This stabilized view can be used for facial expression recognition. We have conducted preliminary testing of detection of eyebrow raises, with encouraging results. Figure 6 shows an example of tracking eyebrow raises. In the future, we plan to utilize this method for facial expression analysis to gain estimates of parameters needed for SignStream annotation, as described earlier.

Hand Shape and Gesture Annotation Algorithms

Another aspect of automatic annotation is estimating the shape (configuration) and 3-D motion of the hands from collected video sequences. Rather than focus on every aspect of the hand gesture annotation problem, we have chosen to focus our effort on the issue of hand shape estimation (in addition to head and facial gesture annotation tools described earlier). At this time, the linguistically significant aspects of hand movement are still not well understood, and, therefore, we believe that development of tools to semi-automate motion annotation would be premature.

To date, many researchers have relied on 2-D tracking in monocular video using blob-based, view-based, or hand contour models. Others have focused on 3-D tracking in multicamera video, with detailed 3-D geometric models of the hand and fingers, and/or recursive estimation of hand orientation/trajectory. For a review, see Pavolovic, Sharma, and Huang (1997). It has been shown that excellent recognition rates can be obtained using only a 2-D blob representation of the hand, given a subset of ASL gestures combined in a constrained syntax. However, in general, detailed hand shape information is necessary to distinguish certain signs from each other. For instance, there are signs that have very similar 2-D blob representations for the hand that can be distinguished only through detailed analysis of information (e.g., palm orientation and finger configurations).

For hand shape estimation, a 2-D view-based representation offers an attractive balance between detailed hand shape modeling, computational speed, and stability. In the view-based representation, hand shape is represented in terms of a collection of example hand images that span the space of possible hand shapes (Poggio & Girosi, 1989; Sclaroff & Pentland, 1994; Ullman & Basri, 1991). In-between and/or novel views can be generated as warps between example hand views, as shown in Figure 7. In our feasibility studies, we employed an image registration formulation that is based on robust statistics (Sclaroff & Isidoro, 1998). The formulation tends to be stable with respect to small self-shadows, some variations in illuminant, and small occlusions.

Another important issue is that of locating the hands and selecting the appropriate hand prototype on the first

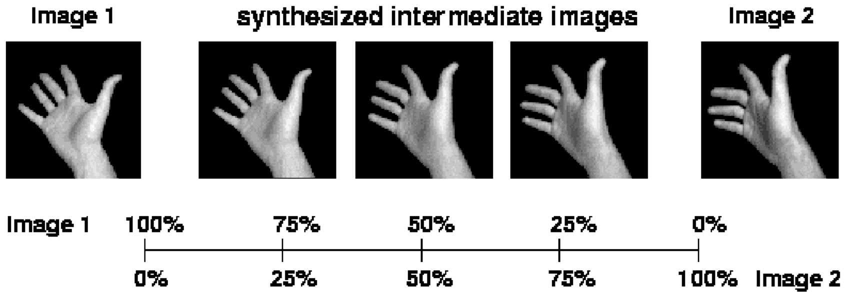


Image 1
synthesized intermediate images
Image 2
Image 1 100% 75% 50% 25% 0%
0% 25% 50% 75% 100% Image 2

Figure 7. Intermediate images are obtained as combinations of the prototype views.

frame of an image sequence. Some existing ASL and gesture recognition systems use automatic skin color and/or motion detection algorithms to detect hands. In SignStream, we are taking a similar approach. However, in ambiguous situations it may be acceptable to ask the annotator to click the mouse on the locations of the hands in the first frame.

The hand tracking system described here is still at the stage of a feasibility study, and integration with SignStream is still a ways off. In the feasibility study, prototype hand images have been selected by a human. In the longer term, there are two issues with regard to prototypes that need to be addressed: (1) algorithms for automatic acquisition of intermediate prototype images and (2) algorithms for eliminating redundancy in prototype space. To address the first issue, the system can automatically acquire new prototypes while tracking; a new prototype should be added if a hand shape differs significantly from those already in the database. For the second, we propose to employ automatic clustering methods to group hand shapes in the prototype collection. Once clusters have been determined, redundancy can be further eliminated via vector quantization methods. Finally, it should be mentioned that methods for incorporating 3-D hand shape are being investigated, since such representations may offer greater accuracy in estimation.

THE NATIONAL CENTER FOR SIGN LANGUAGE AND GESTURE RESOURCES

Through a recent grant from the National Science Foundation, two dedicated facilities for collection of video-based language data have been established, one at Boston University and one at the University of Pennsylvania. Each facility is equipped with multiple synchronized digital cameras to capture different views of the subject. The facilities can capture four simultaneous digital video streams at up to 85 frames per second, while storing the video to disk for editing and annotation. Data collection with the facility began in December 1999. A substantial corpus of ASL video data from native signers is being collected and made available. The instrumentation in the center consists of the following:

1. Four custom built PCs, each with a 500-MHz Pentium III processor, 256 MB RAM, 64 GB of hard drive storage, and Bitflow RoadRunner video capture cards.
2. Four Kodak ES310 digital video cameras. Each camera is connected to one PC.
3. A video sync generator, used to synchronize the cameras. It can be set so that capturing is done in 30, 60, or 85 frames per second.
4. An Ethernet switch, which allows the four PCs to communicate with each other.
5. For synchronized video capture across the four cameras, we use IO Industries' VideoSavant software installed on all PCs.
6. Various illumination sources, black backgrounds, chairs for subjects, and so on.

One of the machines is designated as the "master" machine, and the other three are designated as "slaves." In order to capture a video sequence, we have to start the appropriate program on the master machine, and the appropriate client programs on the slave machines. Then,

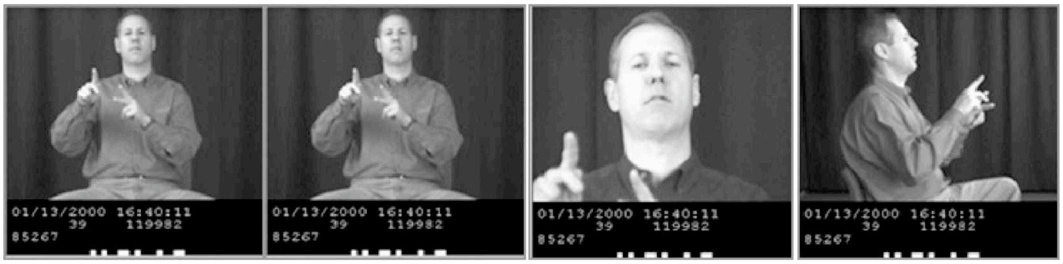


Figure 8. Four example views collected with the instrumentation.

using the master machine, we can specify how many frames we want to capture and start the recording. The captured frames are stored on the hard drives in real time. With 64 GB of hard drive storage available, we can record continuously for 60 min, at 60 frames per second, in all four machines simultaneously, at an image resolution of 648 × 484 (width × height).

In the current setup, sequences have been collected with four video cameras configured in two different ways:

1. All cameras are focused on a single ASL signer. Two cameras make a stereo pair, facing toward the signer and covering the upper half of the signer's body. One camera faces toward the signer and zooms in on the head of the signer. One camera is placed on the side of the viewer and covers the upper half of the signer's body.
2. The cameras are focused on two ASL signers engaged in conversation, facing each other. In this setup, the cameras stand low on tripods placed in the middle (between the two signers, but not obstructing their conversation). One pair of cameras is focused so as to give a close-up facial view of each signer. The other pair of cameras is facing one toward each signer, and covering the upper half of the signer's body.

Example frames from video collected in the facility are shown in Figure 8.

The video data are made available in both uncompressed and compressed formats. Significant portions of the collected data are also being linguistically annotated using SignStream, and these data and the associated SignStream annotations are being made publicly available via the World-Wide Web (see http://www.bu.edu/asllrp/ncslgr.html).

CONCLUSION

One major goal of our research is to provide tools to facilitate detailed and efficient linguistic annotation of visual language data. In this article, we have described SignStream, a database tool that has been designed to facilitate linguistic and computer vision research on signed languages and the gestural component of spoken languages. Despite the advantages provided by tools such as SignStream, computer-assisted manual transcription is still quite time-consuming. Ultimately, automation of aspects of the transcription process will provide the greatest degree of efficiency and accuracy. It is hoped that the use of machine vision algorithms to assist linguists in many aspects of transcription will lead to even greater rapidity in the fine-grained transcription of visual language data, thus further accelerating linguistic and computer science research on sign language and gesture.

The tools that we are developing are being employed in the annotation of a significant new corpus of digital video of native signers of ASL. The video corpus is being collected in the National Center for Sign Language and Gesture Resources, a digital video capture facility at Boston University and at University of Pennsylvania that allows high-resolution digital video capture from up to four simultaneous camera views of the ASL informants.

The availability of large-scale linguistically annotated audio databases spurred linguistic and computational research on spoken language. To date, there are no comparable large-scale corpora for visual language data, even though such corpora are crucial for progress in sign language and gesture recognition. Future computer science research into automatic ASL recognition will require large quantities of data transcribed at a level of granularity that provides significant phonological detail. While it is possible for a transcriber to annotate small samples of data at this level of detail, it is currently not practical to code large quantities of data in this way (at least not in a small number of transcriber-years). In order to overcome this fundamental obstacle to the production of large-scale appropriately annotated corpora, new efficient techniques for annotation must be developed. Ultimately, the best way to accomplish such annotation is through semi-automation (and eventually automation) of aspects of the transcription process via machine vision algorithms. It is hoped that the SignStream system—combined with the machine vision methods for detailed annotation and the type of data collection facility described in this paper—constitutes progress toward the long-term goal.

REFERENCES

Azarbayejani, A., Starner, T., Horowitz, B., & Pentland, A. (1993). Visually controlled graphics. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **15**, 602-605.

Baker, C., & Padden, C. A. (1978). Focusing on the nonmanual components of American Sign Language. In P. Siple (Ed.), *Understanding language through sign language research* (pp. 27-57). New York: Academic Press.

Baker-Shenk, C. (1983). *A micro-analysis of the nonmanual components of questions in American Sign Language.* Unpublished doctoral dissertation, University of California, Berkeley.

Basu, S., Essa, I., & Pentland, A. (1996). Motion regularization for model-based head tracking. *Proceedings of the International Conference on Pattern Recognition*, **1**, 611-616.

Battison, R. (1978). *Lexical borrowing in American Sign Language.* Silver Spring, MD: Linstok Press.

Black, M. J., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. *Proceedings of the International Conference on Computer Vision*, **1**, 374-381.

Brentari, D. (1990). *Theoretical foundations of American Sign Language phonology.* Unpublished doctoral dissertation, University of Chicago.

Brentari, D. (1998). *A prosodic model of sign language phonology.* Cambridge, MA: MIT Press.

Brugman, H., & Kita, S. (1995). Impact of digital video technology on transcription: A case of spontaneous gesture transcription. Kodikas/Code, *Ars Semiotica*, **18**, 95-112.

Crowley, L., & Bérard, F. (1997). Multi-modal tracking of faces for video communications. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition* (pp. 640-645). Los Alamitos, CA: IEEE Computer Society Press.

DeCarlo, D., & Metaxas, D. (1996). The integration of optical flow and deformable models with applications to human face shape and motion estimation. *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, **1**, 231-238.

Essa, I., & Pentland, A. (1997). Coding analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **19**, 757-763.

Hager, G. D., & Belhumeur, P. N. (1998). Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **20**, 1025-1039.

Jebara, T. S., & Pentland, A. (1997). Parametrized structure from motion for 3-D adaptative feedback tracking of faces. *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, **1**, 144-150.

La Cascia, M., Sclaroff, S., & Athitsos, V. (2000). Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3D models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **22**, 322-336.

Li, H., Rovainen, P., & Forcheimer, R. (1993). 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **15**, 545-555.

MacLaughlin, D., Neidle, C., & Greenfield, D. (2000). *SignStream user's guide, Version 2.0* (ASLLRP Report No. 9). Boston: Boston University.

Neidle, C. (in press). SignStream: A database tool for research on visual-gestural language. *Journal of Sign Language & Linguistics*.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., & Lee, R. G. (2000). *The syntax of American Sign Language: Functional categories and hierarchical structure*. Cambridge, MA: MIT Press.

Noldus, L. P. J. J., Trienes, R. J. H., Hendriksen, A. H. M., Jansen, H., & Jansen, R. G. (2000). The Observer Video-Pro: New software for the collection, management, and presentation of time-structured data from videotapes and digital medial files. *Behavior Research Methods, Instruments, & Computers*, **32**, 197-206.

Oliver, N., Pentland, A., & Bérard, F. (1997). Lafter: Lips and face real time tracker. *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, **1**, 123-129.

Pavolovic, V., Sharma, R., & Huang, T. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **19**, 677-695.

Poggio, T., & Girosi, F. (1989). *A theory of networks for approximation and learning* (A.I. Memo No. 110). Cambridge, MA: Artificial Intelligence Lab, Massachusetts Institute of Technology.

Rowley, H. A., Baluga, S., & Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **20**, 23-28.

Sandler, W. (1989). *Phonological representation of the sign: Linearity and nonlinearity in American Sign Language.* Dordrecht: Foris.

Sclaroff, S., & Isidoro, J. (1998). Active blobs. *Proceedings of the International Conference on Computer Vision*, **1**, 1146-1153.

Sclaroff, S., & Pentland, A. (1994, November). *Physically-based combinations of views: Representing rigid and nonrigid motion.* Paper presented at the IEEE Workshop on Nonrigid and Articulate Motion, Austin, TX.

Stokoe, W. C., Casterline, D. C., & Croneberg, C. G. (1965). *A dictionary of American Sign Language on linguistic principles.* Silver Spring, MD: Linstok.

Terzopoulos, D., & Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **15**, 569-579.

Ullman, S., & Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **13**, 992-1005.

Yacoob, Y., & Davis, L. S. (1996). Computing spatio-temporal representations of human faces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **18**, 636-642.

Yuille, A. L., Hallinan, P. W., & Cohen, D. S. (1992). Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, **8**, 99-111.

NOTES

[1] For further information about the syntactic research conducted as part of the American Sign Language Linguistic Research Project, see http://www.bu.edu/asllrp/, which has abstracts of all of our publications and information about many documents and video files available over the Internet or on CD-ROM.

[2] We have had reports of SignStream being used for the study of behavior of patients with different kinds of (linguistic and nonlinguistic) impairments, for example. We have also had inquiries about using SignStream for the study of conducting gestures and animal behavior.

[3] For a discussion of features of SignStream in comparison to SyncWRITER, another commercial application (http://www.sign-lang.uni-hamburg.de/software/syncWRITER/info.english.html) and MediaTagger (described at http://www.mpi.nl/world/tg/CAVA/mt/MTandDB.html and in Brugman & Kita, 1995), see http://www.bu.edu/asllrp/SignStream/comp.html.

[4] Linguistic research on ASL (Battison, 1978; Brentari, 1990, 1998; Sandler, 1989; Stokoe, Casterline, & Croneberg, 1965) has identified four key phonological parameters that define and distinguish ASL signs: hand shape, location (place of articulation), orientation, and movement. Some of these are further decomposed into several features; for example, hand shape is modeled by Brentari in terms of significant fingers, aperture (e.g., open or closed), and joints (bending at the base or nonbase joints). Differences in these features can distinguish different hand shapes.

[5] There is an extensive literature on many of the linguistic functions of nonmanual gestures; see, for example, Baker and Padden (1978), Baker-Shenk (1983), or Neidle et al. (2000) for a summary and further references. The essential role of nonmanual gestures in the grammars of signed languages is indisputable. However, it is also the case that there are nonmanual markings whose linguistic function is not yet fully understood. Tools of the kind that are being developed as part of this project will also be invaluable in further linguistic study of linguistically significant nonmanual gestures.

(Manuscript received September 28, 2000;
revision accepted for publication May 10, 2001.)