# Intelligent Temporal Subsampling of American Sign Language Using Event Boundaries

David H. Parish, George Sperling, and Michael S. Landy
New York University

How well can a sequence of frames be represented by a subset of the frames? Video sequences of American Sign Language (ASL) were investigated in two modes: dynamic (ordinary video) and static (frames printed side by side on the display). An activity index was used to choose critical frames at event boundaries, times when the difference between successive frames is at a local minimum. Sign intelligibility was measured for 32 experienced ASL signers who viewed individual signs. For full gray-scale dynamic signs activity-index subsampling yielded sequences that were significantly more intelligible than when every $m$th frame was chosen. This result was even more pronounced for static images. For binary images, the relative advantage of activity subsampling was smaller. We conclude that event boundaries can be defined computationally and that subsampling from event boundaries is better than choosing at regular intervals.

American Sign Language (ASL) is a gestural form of communication used by the North American deaf and hearing-impaired communities. In free conversation, ASL is as rapid a form of communication as most spoken languages, including English (Bellugi & Fischer, 1972). Over the past decade there have been several investigations of factors related to the transmission of ASL over the existing long-distance communications networks. The problem is to compress a video signal of the signer to the extent that it will fit through a low-bandwidth or low bit-rate communication channel, such as an ordinary telephone line, without greatly disrupting the efficiency of communication. Although previously designed video telephones would suffice for communication, their bandwidth requirements and cost made them impractical. The current public telephone network has transmission limits of 300 to 2800 Hz for analog signals and nominally 9,600 bits per second (bps) for digital signals (ca. 1988).

## Methods of ASL Compression

### Spatial Compression

The problem of determining the minimum communication requirements for ASL was posed by Sperling (1978), who subsequently determined that remarkably sparse images could convey intelligible messages (Sperling, 1980, 1981). Sperling, Landy, Cohen, and Pavel (1985) investigated ASL intelligibility versus bandwidth trade-offs by using the most sophisticated spatial image-compression and coding schemes then available. In one condition, subjects were able to interpret signs with normalized intelligibility of .86, in relation to full-bandwidth sequences, even though bandwidth had been reduced to 2880 Hz. Related work by Abramatic, Letellier, and Nadler (1982) with French Sign Language and by Pearson (1981) with British Sign Language also offers the possibility of substantial compression.

The relative success in ASL compression achieved by Sperling et al. (1985) may be attributed to the large amount of redundant spatial information within the ASL signal. Spatial redundancy exists both across individual pixels and across groups of pixels. Individual pixels are redundant when the gray level of one pixel is predictive of the gray level of nearby pixels. Groups of pixels are redundant when cues in one region of the image yield predictions of what should appear in other regions. For example, consider the particular configuration of pixels that yields the form of an arm. To the degree that the hand, elbow, and shoulder are resolvable, other arm pixels are probably unnecessary and are therefore redundant. Accordingly, one may discard some forms of spatial information with the expectation that other, similar information remains. This sort of spatial redundancy provides the basis for the often surprising success of dynamic point-light displays (Cutting, 1978; Johansson, 1973) in conveying rather complex form information. Indeed, Poizner, Bellugi, and Lutes-Driscoll (1981) demonstrated that such displays successfully con-

vey important lexical and inflectional information in ASL and that subjects are quite adept at identifying signs when so presented.

## Temporal Compression

In the present study, following Pearson (1981) and Sperling et al. (1985), we investigate temporal redundancy in ASL. By identifying redundant dynamic information, we are able to intelligently subsample an ASL signal in time and thereby retain intelligibility while decreasing the number of transmitted frames. What criteria should guide temporal subsampling? We consider an ASL sequence as a series of events that unfold over time. The subsampled frames should be those that best convey these events. For example, an event might involve moving the left arm from a position parallel to the body to a position perpendicular to the body. It is important to keep clear the distinction between motion and events. *Motions* are continuous and occur at one or more locations simultaneously. An *event* is a connected sequence of frames, the result of segmenting an image sequence. Just as an object in the *x, y* domain is a grouping together of pixels to which a common label will be applied, an event is a grouping of frames under a common label.

In some instances, the perception of events requires a cognitive component, an interpretation by the viewer (Ebbesen, 1980; Markus & Zajonc, 1985). In other instances, however, there is strong support for the theory that events (or perceptual units) are directly perceived much like motion (Asch, 1952; Heider, 1958; Newtson, Hairfield, Bloomingdale, & Cutino, 1987). Insofar as events can be perceived directly, it implies that there must be consistent stimulus properties that reflect the event structure within a motion sequence.

How does one locate and use such stimulus properties for temporal compression? Interestingly, the problem of dynamic sequence segmentation has been addressed in several domains of study and for a variety of purposes. These include efforts to construct humanlike motion representation schemes (Marr & Vaina, 1980; Rubin & Richards, 1985), the development of motion descriptors for robotics or artificial intelligence (Thibadeau, 1983), and, as noted previously, cause attribution and event perception in the field of social psychology (Heider, 1958; Newtson, 1973). In another ASL study, Green (1984) attempted to locate the boundaries between consecutive signs in a stream of ASL images.

## Perceptual Units of Behavior

A technique for determining the location of boundaries of perceptual units that has received considerable attention comes from Newtson (1973) and his collaborators (Newtson & Engquist, 1976; Newtson, Engquist, & Bois, 1977; Newtson & Rindner, 1979; Rindner, 1982). Although Newtson's early work was concerned with determining how the attribution process varies as a function of the unit of perception, he was also concerned with demonstrating that there was, in fact, an objective basis for determining units of behavior. In his pro-

cedure, perceptual units were marked by subjects who viewed films and pressed a button connected to a continuous event recorder. Subjects had been instructed to press the button when, "in your judgment, one meaningful action ends and a different one begins" (Newtson, 1973, p. 30). There was a high degree of consistency, both within and between subjects, as to where the boundaries were placed.

In a series of experiments, Newtson and Engquist (1976) showed that subjects were more sensitive to disruptions at the boundaries of behavioral units (breakpoints) than at moments between the boundaries (nonbreakpoints), suggesting that the breakpoints had a high degree of psychological salience. They noted that deletions of frames from ongoing films at breakpoints were more accurately detected than deletions of frames at nonbreakpoints. Moreover, subjects who viewed static presentations of three consecutive breakpoints were more accurate in their descriptions of the action, rated the sequence as more intelligible, and were more accurate at ordering the slides when the slides were shown out of order than subjects who were shown other groups of three frames. In one experiment (Newtson & Engquist, 1976), groups of three breakpoint frames were rated as high on intelligibility as the continuous sequences from which they were drawn. In short, Newtson and his collaborators provided a convincing demonstration that subjectively defined breakpoint frames convey information that is of greater importance to the global percept than nonbreakpoints. (For a review of the role of breakpoints in event perception, see Newtson et al., 1987).

## Choosing Significant Frames Automatically

In the present study, we measure the intelligibility of ASL sequences constructed by using a subset of chosen frames from the original sequence. The work of Newtson and his collaborators suggests that in segmenting the sequences, we would do best to retain breakpoints while discarding nonbreakpoints. Although the Newtson procedure for locating breakpoints works well, its usefulness in a real-time communication system is limited, because human observers must select the frames. Given the growing availability of digital image-processing technology, it is natural to digitize the image sequences to be transmitted, to compute which frames represent breakpoints within the sequence, and to transmit the chosen frames and discard the remainder. To implement such a system, we must first consider the physical properties associated with breakpoints: the boundaries of perceptual units.

## Physical Characteristics of Breakpoints

Two possible theories could account for the finding that an action stream may be partitioned into discrete units of behavior based on breakpoints (Newtson & Engquist, 1976). Either the particular configuration of components in the scene constitutes a *distinctive state* that identifies the breakpoint, or actions are defined by *state-to-state* changes that are characterized by successive breakpoints. In a test of these two possibilities, Newtson et al. (1977) used a movement notation system, designed for use by choreographers, to code the body

positions of the actors in their sequences. Differences between codings at different points in time describe the position changes of the actor; the surface structure of each sequence is captured. By comparing position changes between successive breakpoints, between breakpoints and nonbreakpoints, and between randomly chosen nonbreakpoints, Newtson et al.'s (1977) results demonstrated strong support for the state-to-state hypothesis but showed no evidence for the distinctive state hypothesis.

The results of these experiments are interpreted to mean that the actors' positions were maximally distinct, within the constraints of the behavior, at successive breakpoints (Newtson et al., 1977, 1987). To achieve maximally distinct body positions at breakpoints, some form of activity must have occurred between breakpoints; that is, the position change cannot happen instantaneously. Newtson et al. (1977, 1987) interpreted their measure of position change over time as an index of movement complexity. This measure, as they pointed out, is related to, but different from, movement magnitude. The important fact that we derive from this work is that some form of dynamic activity must occur between successive breakpoints. This empirical observation supports our subjective impression that some greater-than-average amount of activity occurs between boundaries and that the activity level decreases at boundaries.

Evidence of such activity must be available in the surface characteristics of the sequence. Marr and Vaina (1980) formalized this notion in their state–motion–state representation for segmenting a stream of movement into pieces that can be described independently. They used pauses (described as moments when the parts of a shape are either absolutely or relatively at rest) to segment a motion stream. Pauses occur when the object (or objects) in the scene undergo a change in direction of movement and occasionally occur at other moments in a sequence. This same notion appears in the work of Rubin and Richards (1985), who argued that natural motion boundaries occur at starts, stops, and force discontinuities. Moreover, they provided evidence that human observers have a subjective impression that a significant event has occurred at each of these boundaries. All of these theories imply that one ought to be able to locate event boundaries by tracking some surface characteristic of the dynamic sequence and by searching for frames that correspond to pauses in activity. Rubin and Richards (1985) and Marr and Vaina (1980) theoretically defined ways in which motion sequences might be parsed. In this article we choose a simple realization of these ideas and test its effectiveness for producing intelligible subsampled motion displays.

## The Activity Index, $a_\theta(n)$

The activity index is the fraction of pixels that experience a suprathreshold change in luminance between frames $n - 1$ and $n$. We located event boundaries in ASL sequences by computing this measure of activity between each pair of consecutive frames in each sequence and looking for the local minima. Our activity index was computed by counting the number of pixels that underwent a significant change of gray level between consecutive frames in a sequence; the fraction

of such pixels is reported for each frame after the first. This scheme takes advantage of the fact that the percept of motion results from the changing of luminance values over space and time; more activity causes greater numbers of pixels to change from frame to frame. Obviously, it works best when the camera and luminance sources are stable. In a typical ASL sequence, several body parts move simultaneously. Thus, $a_\theta(n)$ is more indicative of the general level of activity throughout the sequence than of the motion of any particular object.

A sequence consists of $N$ frames; $X$ is the number of rows and $Y$ is the number of columns in each frame. The luminance value at spatial location $x$, $y$ in frame $n$, $1 \leq n \leq N$ is $l(x, y, l, n)$. The number of pixels that change in luminance by more than a threshold amount $\theta$ is computed by setting

$$t_\theta(x, y, n) = \begin{cases} 1 & \text{if } |l(x, y, n) - l(x, y, n - 1)| > \theta \\ 0 & \text{otherwise.} \end{cases}$$

The activity index $a_\theta(n)$ is the fraction of pixels in frame $n$ that experienced a suprathreshold change in luminance:

$$a_\theta(n) = \frac{\sum_{x=1}^{X} \sum_{y=1}^{Y} t_\theta(x, y, n)}{XY}.$$

The higher the threshold parameter $\theta$, the smaller the influence of pixels that change as a result of camera or digitizing noise rather than as the result of a moving object.

As an example, we consider a 30-frame sequence in which a white square on a black background moves left and right across the width of a frame sinusoidally. Figure 1 is a graph of $a_\theta(n)$ for such a sequence. Note that the local minima in Figure 1, where $a_\theta(n - 1) > a_\theta(n) < a_\theta(n + 1)$, correspond to frames in the original sequence in which the direction of motion is changing (i.e., the peaks and troughs of the sinusoid). Complex movies such as ASL sequences produce complex $a_\theta(n)$ functions with many more local minima.

Figure 2 shows the $a_\theta(n)$ function for a 70-frame ASL sequence that shows the sign for the word accident. A drawing
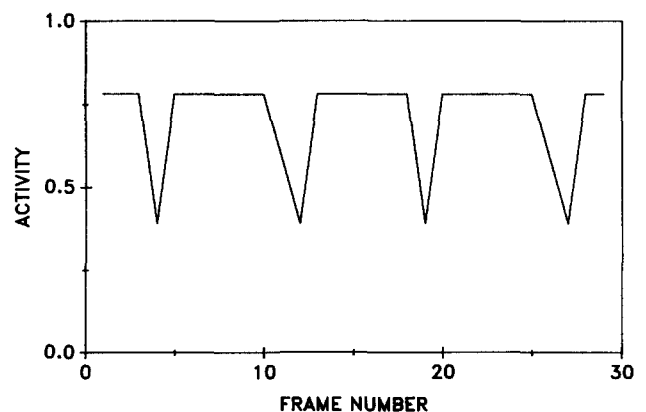


Figure 1. Activity index as a function of frame number for a small square that moves through two cycles of a sine wave across 30 frames. (The local minima correspond to the resting points, or points of change in direction, of the moving object.)
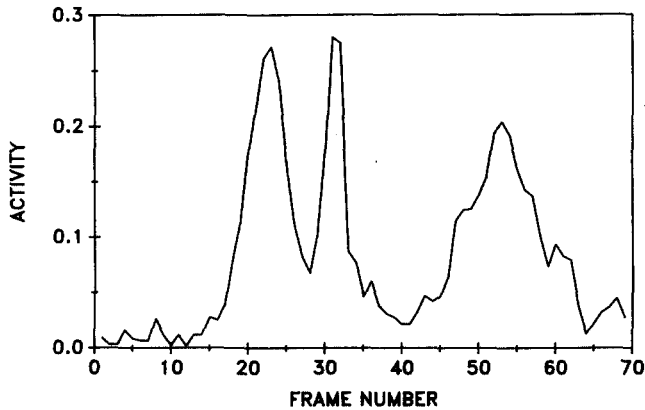
*Figure 2.* Activity index as a function of frame for a 70-frame ASL sequence that shows the sign for the English word *accident*. (The large amounts of activity at the beginning and end of the function correspond to the action of the signer as she moves into and out of a rest position [arms folded in front]. The two local minima at Frames 28 and 40 correspond to moments when the signer's hands are furthest apart and when they meet in the middle.)

The two large local minima that occur at Frames 28 and 40 in Figure 2 correspond to the point in the sign when the signer's hands are spread apart and to the frame in which they meet. In other words, if we choose frames that correspond to the two local minima at Frames 28 and 40, as well as a beginning and ending frame, we satisfy the criteria for intelligent temporal sampling while reducing the sequence from 70 to 4 frames. These 4 frames are illustrated in Figure 5a.

*Rotational and circular motions.* Because the activity index relies on changes in activity to indicate event boundaries, rotational or smooth circular motion presents a problem: An activity index may not achieve a significant local minimum during such a motion. Marr and Vaina (1980) recognized the same shortcoming in their state–motion–state representation and proposed to handle these instances by recognizing the occurrence of a confounding movement and dealing with it separately. Our $a_\theta(n)$ activity index treats rotational and circular motions the same as all others. If we were to discover that subsampled signs never reached some minimum criterion of intelligibility, it might indicate that rotational and circular motion occur often enough within ASL to merit special consideration. However, an informal survey of signs suggests otherwise.
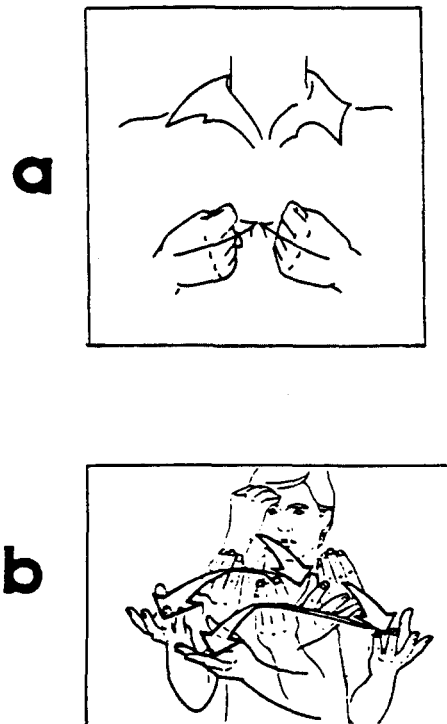
of the sign is given in Figure 3a, and every third frame is shown in Figure 4. In our experiment, the signer assumes a rest position in which her arms are folded in front of her at the beginning and end of the sequence. The same rest position is used for all signs in order to remove any potential ambiguity concerning the beginning and ending of each sign and to ensure that on repeated presentations, the particular frame on which a sign begins or ends does not provide a clue to the identity of the sign. In the sign that produced Figure 2, the signer raises her two hands to either side, closes them into fists, and moves them until they meet in front of her (this is an iconic sign for a collision), and then reassumes the rest position with arms folded.

The $a_\theta(n)$ function in Figure 2 is instructive for several reasons. First, there is always a high activity value at the beginning and end of each sequence as the signer moves out of and into the rest position. In Figure 2, these peaks occur at Frames 23 and 53. Moreover, at the beginning and end of the sign, the activity index becomes a collection of closely spaced local minima. These frames correspond to the rest position in which luminance noise and slight movements on the part of the signer account for fluctuations in $a_\theta(n)$, and which might be misinterpreted as significant activities (i.e., these frames might be selected).

*Activity-index subsampling.* Activity-index subsampling means selecting for presentation only the frames for which $a_\theta(n)$ has a relative minimum as a function of $n$. To control the coarseness with which candidate minima are sampled, we introduced a parameter $\alpha$ that specifies the minimum increase in $a_\theta(n)$ that must occur between consecutively chosen frames; that is, in order to choose both frames $f_i$ and $f_j$, where $i < j$, the activity index must rise above $a_\theta(i) + \alpha$ for some frame $k$, where $i < k < j$. Note that this method of sampling is asymmetric with respect to time; inverting the order of frames may lead to a different selection.



*Figure 3.* Two illustrated signs showing (a) a simple sign ACCIDENT (from *The Joy of Signing* [p. 102] by L. L. Riekehof, 1980, Springfield, MO: Gospel Publishing House. Copyright 1980 by Gospel Clearing House. Reprinted by permission) and (b) a compound sign I-INFORM-YOU-TWO (from *A Basic Course in American Sign Language* [p. 158] by T. Humphries, C. Padden, and T. J. O'Rourke, 1980, Silver Spring, MD: T. J. Publishers. Copyright 1980 by T. J. Publishers. Reprinted by permission).
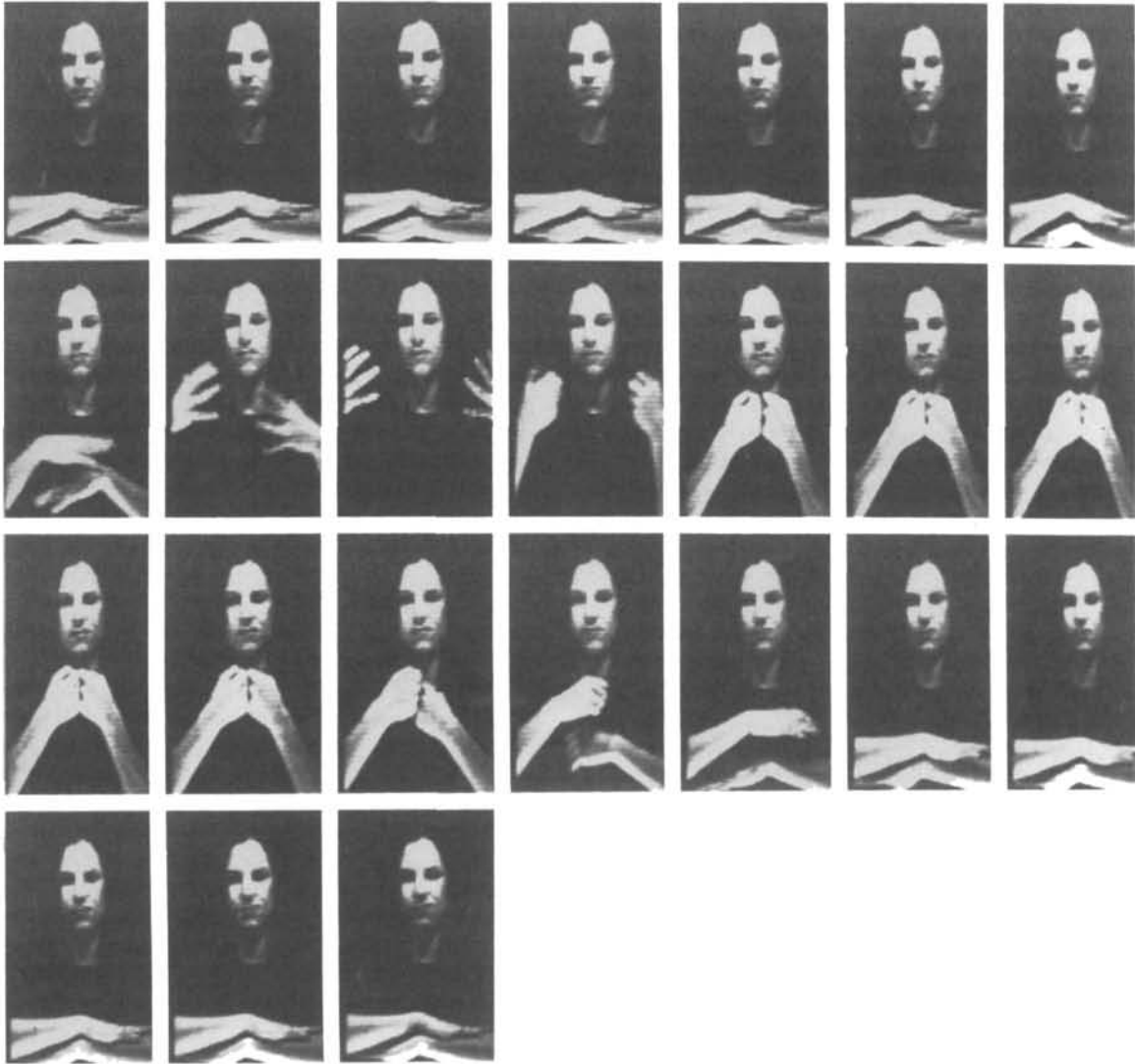
*Figure 4.*   Sequence of digitized images of the sign *accident.* (Every third frame of a 70-frame sequence is shown. This is constant subsampling.)

## Constant Subsampling

To measure the use of activity-index subsampling, it must be compared with an alternative method of temporal compression. Although the focus of their work was spatial rather than temporal compression, Sperling et al. (1985) and Pearson (1981) used a simple frame repetition, what we here call *constant temporal subsampling.* By this method, every *m*th frame is chosen from the sequence, where *m* can take on any value between 2 and the total number of frames in the original sequence. Constant subsampling will be used as the basis of comparison in the present study. Figures 4 and 5b illustrate constant subsampling for *m* equal to 3 and 23, respectively. In Figure 5b, note that the second frame catches the signer in the middle of a movement.

## Dynamic Display Considerations

Having chosen a subset of the frames from a sequence of *N* frames, how should we choose the duration of each frame to ensure that the displayed sequence retains as much of the rhythmic properties of the original as possible? Temporal constancy is preserved in constant subsampling by choosing the number of repetitions for each frame equal to the constant sampling factor. For example, if every third frame were chosen from the original sequence, each frame in the displayed sequence would be repeated three times.

Because the frames chosen from a complex scene via the activity index are not necessarily separated by a constant number of frames, the repetition factor for display must vary according to the location of the chosen frame in the original
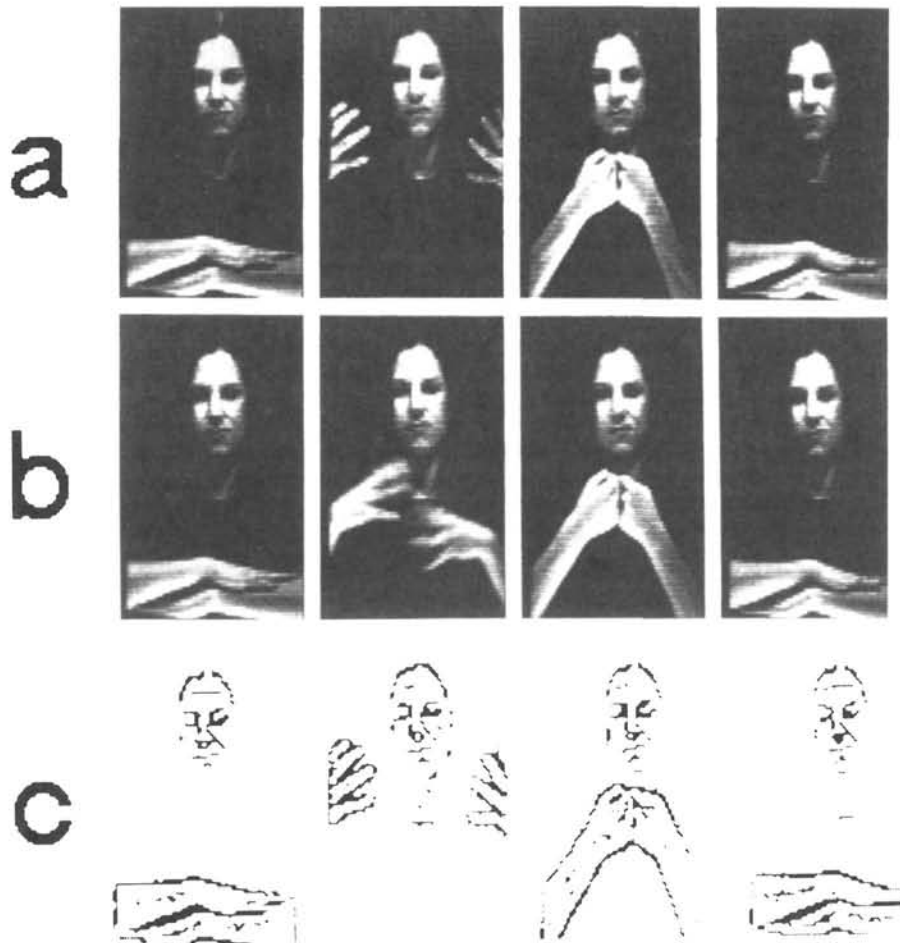
*Figure 5.* Four-frame representations of the sign *accident*: (a) full gray-scale using activity-index subsampling, (b) full gray scale using constant subsampling, and (c) binary images using activity-index subsampling.

sequence. We repeat each chosen frame (to replace discarded frames) until the next chosen frame occurs. For example, if Frames 1, 5, 15, and 22 were chosen from a 30-frame sequence, Frame 1 is repeated 4 times, Frame 5 is repeated 10 times, Frame 15 is repeated 7 times, and Frame 22 is repeated 9 times, to reach the total of 30 frames that appear in the original sequence. In this method, a different display sequence would be produced from the same sequence of selected frames when played in the forward rather than the time-reversed direction.

## Static Presentation

### *Optimal Number of Frames*

ASL-related investigations have, to this point, focused exclusively on the transmission and intelligibility of dynamic images. There are, however, several compelling reasons for studying the intelligibility of ASL when it is presented in static form. Most important, static images are used in most, if not all, ASL textbooks and dictionaries (e.g., Humphries, Padden, & O'Rourke, 1980; Riekehof, 1980). Important exceptions are the books produced by Stokoe and his collaborators (Stokoe, 1974; Stokoe, Casterline, & Croneberg, 1976), which use written symbolic notation to convey the motion and hand shape of each sign.

The type of static presentation that is most often seen in standard ASL textbooks is a single-frame image that corresponds roughly to a single English word or expression. Typically, an illustrated signer is presented with overlaid arrows and "strobe" lines to indicate the desired hand, finger, and arm motions. An example of one such illustration, from Riekehof (1980), appears in Figure 3a. For simple signs, especially those that use only one hand, these illustrations are quite efficient. Difficulties can arise, however, for compound signs that require a change in hand shape or for the occasional presentation of complete sentences. In these instances, such as for the sign depicted in Figure 3b (taken from Humphries et al., 1980), the many strobe lines and arrows mask the

intended movement and make it difficult for students to replicate the sign.

An alternative to presenting a single frame for each English word or phrase is to present the frames arranged adjacently, as in comic strip format (see Figures 4 and 5). Given the impracticality of displaying the hundreds of frames that may constitute a single sentence, it is necessary to choose a subset of frames for presentation. How does one choose frames in order to convey a sign? Obviously, this is the static analog to the dynamic display that has been addressed earlier and, coincidentally, is a question of great importance to animators and cartoonists. When depicting an action sequence, animators are taught to represent the extremes of the activity first, and then to fill in with in-between frames as needed (Levitan, 1960). In the context of ASL, if frames are chosen to successfully convey the motion when displayed dynamically, do these frames convey the same information when displayed statically?

## Spatial–Temporal Compression Trade-Off

Finally, it is useful to investigate interactions between spatial and temporal compression. A practical application would probably combine temporal and spatial compression in order to avoid the degrading effects of removing too much of either spatial or temporal information. Here, we measure intelligibility for both full gray-scale ASL sequences (8 bits/pixel) and for the same sequences made binary with an edge detection scheme. Each image is convolved with a Gaussian-smoothed Laplacian and then thresholded so that 10% of the values are set to black, generally from the dark side of image edges. The result is a binary, line-drawn image (with approximately 0.211 bits/pixel). An example of such a binary sequence is shown in Figure 5c. (These sequences and the nominal data rate were taken from Sperling et al., 1985, Experiment 2, Condition H.)

## Method

### Subjects

The 32 subjects used in this study were recruited in various places, including the New York Society for the Deaf, the New York Univer-

sity Office for Disabled Students, and word-of-hand among the deaf community. Several fluent hearing ASL interpreters were also used. The mean age of our subjects was 33 years (ages ranged from 18 to 52), and they had been signing for an average of 18 years. Twelve native signers—those who were raised in homes where ASL was the primary language—were included in the study.

### Stimuli

The stimulus set consisted of 84 ASL signs, each of which corresponds roughly to a single English word. All signs were taken from Sperling et al. (1985), who filmed, digitized, and applied various image transformations to the signs for use in their study of trade-offs between ASL sign intelligibility and bandwidth. The signer was filmed from approximately 10 ft (approximately 3.05 m) away, so that the upper body and head filled the viewfinder of the camera. During filming, the signer stood behind a screen with a 12 × 18 in. aperture, wore dark clothing, and had dark hair; these conditions ensured that the hand and face of the signer would be highlighted. Each digitized frame was subsequently cropped to 96 × 64 pixels; the signer was centered in each frame so that the area from her waist to the top of her head was visible.

Along with the original full gray-scale (denoted FGS) movies of each sign, we used signs that had been transformed from the FGS to the line-drawn, binary images previously described. Such signs will be referred to as BIN (for binary) signs. Sperling et al. (1985) reported an intelligibility of .911 for BIN signs, normalized against the percentage correct for 96 × 64 pixel FGS signs. At the time of the experiment, four FGS signs were not available under the BIN image transformation; the BIN conditions used four signs that did not appear in FGS conditions, and vice versa. The list of 84 signs used in the study appears in Table 1, divided into the stimulus blocks used in the experiment.

Although the term *frames per second* (fps) is used throughout the remainder of this article to describe the degree of temporal compression, all dynamic stimuli were presented on a system that always displayed 60 fps. In the context of the present study, fps refers to the number of new frames per second, computed by dividing the number of chosen frames by the duration of the original (or the derived) sequence. Frame rate (fps) was varied parametrically. The parameters $m$ and $\alpha$ control the frame rate for fixed-rate and activity-index subsampling, respectively. Four different values were used for each of these two parameters. This manipulation allows us to collect data over a large range of intelligibility. The parameter values and the average number of frames per second for each scheme is displayed in Table 2.

Table 1
*Stimulus Blocks*

| Block 1 | Block 2 | Block 3 | Block 4 | Block 5 | Block 6 | Block 7 | Block 8 |
|---|---|---|---|---|---|---|---|
| telegraph | wrong | sit | emphasize | punishment | bear | tobacco | summer |
| leave | general | cheese | wife | apple | kill | thanks | think |
| deaf | girl | until | world | uncle | flag | home[a] | talk |
| finish | short | shoe | our | screwdriver | sorry | flower | wrestling |
| plan | week | wait | accident | guilty | tree | member | love |
| ugly | noon | picture | hospital | paper | bread | challenge | read |
| train | preach | month | friday | understand | behind | pay | start |
| relax | red | steal | cancel | yesterday | everyday[a] | fun | color |
| mother | machine | program[a] | because | letter | eye | which | before |
| jump | improve | spend | boss | bored | cop | grow | alive[a] |
| | | egg[b] | | | movie[b] | pour[b] | lousy[b] |

[a] Signs that appeared only in BIN (binary) conditions.  [b] Signs that appeared only in FGS (full-gray scale) conditions.

Table 2
*Stimulus Transformations and Frame-Rates*

| Scheme | $\alpha$ or $m$ | Frames per second |
|---|---|---|
| Activity index | 0 | 10.8 |
| Activity index | 0.02 | 8.85 |
| Activity index | 0.05 | 6.75 |
| Activity index | 0.1 | 5.4 |
| Constant | 2 | 30 |
| Constant | 4 | 15 |
| Constant | 7 | 8.6 |
| Constant | 11 | 5.5 |

*Note.* The $\alpha$ is the parameter that controls the number of samples used by the activity-index sampling scheme; $m$ is the parameter that governs the number of frames chosen by the constant subsampling scheme.

## Procedure

The ASL signs were divided into eight groups of 10 signs, each group balanced for difficulty by the criterion of Sperling et al. (1985). The experimental variables included two image types FGS and BIN; two presentation modes, dynamic (D) and static (S); two subsampling schemes, constant and activity index; four frame rates; and 10 stimulus blocks. A full-factorial experiment on these factors would require 320 subjects with only 1 subject in each cell. To achieve a more manageable study, we ran four separate groups of subjects, one for each combination of image transformation and presentation mode (FGS-D, FGS-S, BIN-D, and BIN-S).

To make the most efficient use of each subject, the remaining factors within each of the four groups were subjected to a Greco-Latin design in which subsampling scheme and compression factor were fully randomized and order of presentation was partially randomized. In other words, each stimulus block of 10 signs was paired one time with every combination of subsampling scheme and compression factor over the course of the experiment. For convenience, the combination of subsampling scheme and compression factor is referred to as the stimulus *transformation*; every transformation appeared in each ordinal position of stimulus presentation. Order of presentation is only partially randomized, because sequence effects are not balanced in this design. Each subject saw eight complete stimulus blocks, each block having undergone a different transformation (i.e., repeated measures over transformation and stimulus block). A total of 32 subjects were required for a single replication through each of the four 8 × 8 Greco-Latin squares.

*Intelligibility test.* All stimuli were processed with the HIPS image-processing software (Landy, Cohen, & Sperling, 1984a, 1984b) and were presented on a computer-controlled graphics display processor (Adage RDS-3000 image-processing system). Images were viewed on a Conrac 7211C19 monitor, set so that the mean luminance of the display was equal to 55 candela per square meter (cd/m²). Subjects were seated approximately 1 m from the screen, though they were free to move to their most comfortable distance. (Parish & Sperling, 1987, demonstrated that for stimuli whose visibility is impaired by noise, viewing distance, over an extremely wide range, is immaterial.)

For all conditions, subjects were required to respond to each ASL presentation with an English gloss for the presented sign. Subjects were told that each sequence contained only a single sign and that each sign corresponded, roughly, to one English word. In most cases, subjects wrote their responses on an answer sheet. In cases in which deaf signers did not possess English skills that were advanced enough to allow them to respond with a written word, they would sign the response to an ASL interpreter who then recorded the English equiv-

alent. The interpreter confirmed these subjects' understanding of the sign by having them either use the word in a sentence or further elaborate on the meaning of the word. Finally, all subjects were told that if they had no idea what the correct answer was, they did not have to respond.

*Dynamic presentation.* The word *begin* appeared on the monitor, signaling the subject to press any button on a five-button keypad. After the button press, the screen was cleared, and a white cue spot appeared for 0.5 s. This was followed by a 0.5-s blank interval and the presentation of an ASL movie (frame sequence). The sequence was shown once, with the frames repeated as necessary to retain the duration of the original image sequence, and was followed by a blank screen. The word *wait* was displayed until the next sequence was ready for display (2 or 3 s), at which point the word *continue* appeared. While waiting for *continue* to appear, we recorded the subject's response. After the subject's response and after the word *continue* appeared on the screen, the subject was free to press any button to initiate the next trial.

*Static presentation.* As with the dynamic presentation, the initial button press erased the word *begin* from the screen and caused the stimulus to be presented, though without a cue spot. The frames of each movie were arranged in order by rows and columns, from left to right and from top to bottom. Up to seven frames appeared in each row. A sample "page" of 24 frames for the sign *accident* is shown in Figure 4. Shorter pages are shown in Figure 5 for several conditions. On presentation, the subject scanned the page and decided on a response. Before writing or signing the response, however, a second button press was required to erase the screen. After responding, the word *continue* appeared on the screen. The next button press initiated the next trial.

## Results

### Scoring

The measure of performance for all subjects and conditions is in percentage correct. For some of the signs used in the study, several English responses are considered correct, a result of the historical and regional development of ASL. Each subject's answer sheet was scored by a congenitally deaf signer who is fluent in ASL.

### Subject Comparison

To assess the general ability of the subjects in this study, we may compare their performance for dynamic sequences with the performance of the subjects from Sperling et al. (1985), who viewed similar sequences. For the most richly subsampled full gray-scale sequences, which averaged about 20 frames, subjects in the present study averaged 86% correct, nearly identical to the Sperling et al. subjects, who averaged about 87% correct. For binary images, subjects in the present study averaged 70%, in comparison with Sperling et al.'s 80% correct. Our subjects may have been somewhat less skilled than those of Sperling et al., perhaps a result of the mix of native and nonnative, hearing and nonhearing signers used in the present study, as opposed to the more homogeneous group of deaf signers used by Sperling et al.

### Main Effects

The data in each of the four Greco-Latin squares, distinguished by the combination of image type and presentation
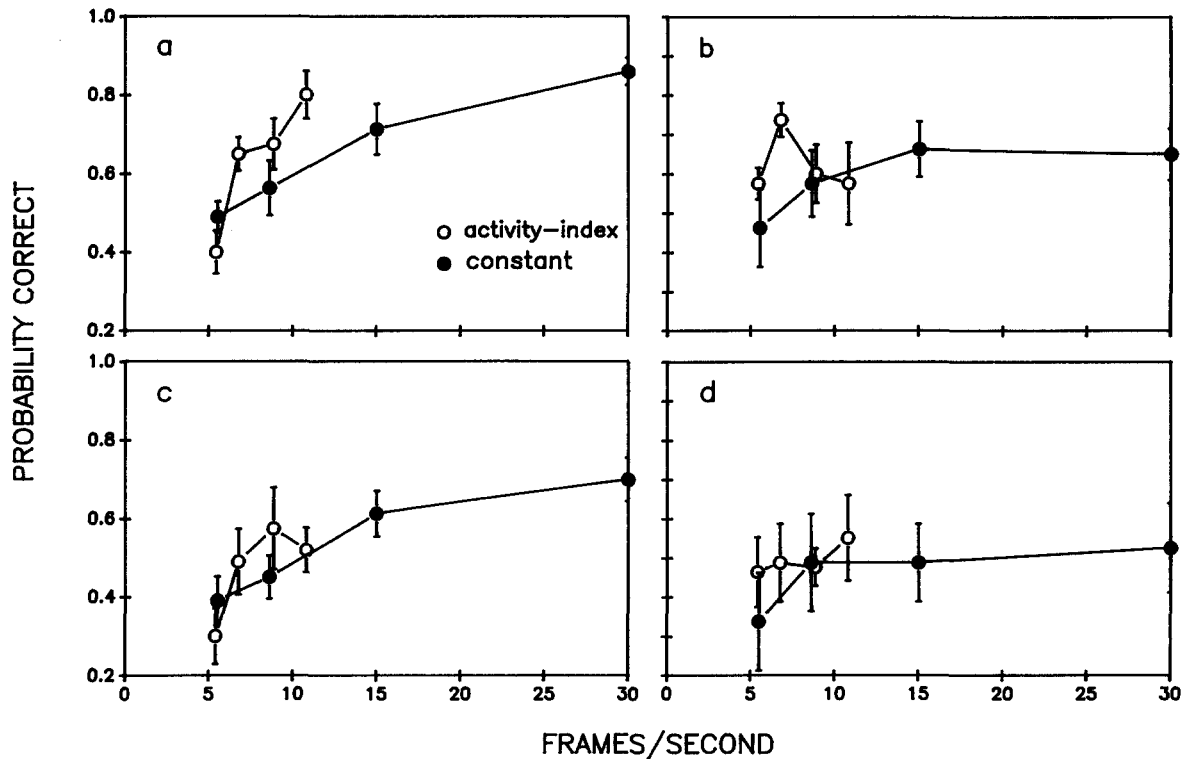
*Figure 6.*  Mean subject performance as a function of the mean number of frames per second for each of the four main conditions of the experiment and for each subsampling scheme: (a) full gray-scale images in dynamic presentation (FGS-D), (b) full gray-scale images in static presentation (FGS-S), (c) binary images in dynamic presentation (BIN-D), and (d) binary images in static presentation (BIN-S). (The open dots on each graph represent performance with activity-index sequences and the solid dots represent constant subsampled sequences. The vertical bars represent the standard error of the mean.)

mode, are displayed in Figure 6. Probability correct is displayed as a function of mean number of frames, averaged across the 8 subjects within each design. These data were subjected to an arcsine transformation in order to decorrelate mean and variance; the arcsine data were used in the subsequent analyses. Main effects for each individual Greco-Latin square were evaluated by an analysis of variance.

The four main effects for each Greco-Latin square are subjects, order, stimulus block, and image transformation. The most stringent assumption made by the analysis is that the interactions among the four main effects of the Greco-Latin square are negligible (Winer, 1971). Every effort was made to ensure negligible interactions: Subjects were assigned randomly to each cell, and stimulus blocks were balanced for difficulty. There is no a priori reason to assume that there would be significant interactions.

Stimulus transformation, which includes both subsampling scheme and compression factor, was a significant factor for the FGS-D, BIN-D ($p < .01$), and FGS-S ($p < .05$) conditions but not for the BIN-S condition, although there was a trend in the expected direction. Stimulus block was significant for all four designs ($p < .05$). The fact that the stimulus blocks differed from each other indicates that our efforts to equate the blocks for difficulty was not entirely successful. This is not surprising, because Sperling et al. (1985) also found a

significant effect of stimulus block despite similar efforts. We rely on the fact that throughout the course of the experiment, all stimulus blocks were presented in all conditions, thereby allowing block effects to balance out. Finally, there were significant subject differences ($p < .01$) for static presentation of both image types (FGS and BIN).

## Subsampling Scheme

The data from the full gray-scale conditions, seen in Figures 6a and 6b, suggest that the activity-index sequences were more intelligible than constant subsampled sequences. Ideally, we would have had data from both schemes at the same frame rate to allow us to directly test this hypothesis. Unfortunately, the nature of the subsampling schemes prevents such sampling precision. To conduct the test, we used linear interpolation to estimate performance for 6.75 fps for constant subsampling. Activity-index data had already been collected at this frame rate. For each presentation format, a $t$ test was computed with the interpolated constant-subsampling data and real activity-index data. Both tests strongly reject the null hypothesis ($p < .01$). For dynamic presentation of binary images (Figure 6c), data interpolated at 8.85 fps for constant subsampling also reject the null hypothesis ($p < .05$).

For dynamic presentation, activity-index performance is estimated, by averaging between points, to be about 8% better than for constant subsampling (for the portions of the curves that overlap). This estimate reflects, in part, the crossover interaction that occurs at the lowest frame rate, in which constant subsampling outperforms activity-index performance. This crossover interaction almost certainly comes from the fact that the activity-index scheme chose frames nearer to the beginning and end of the original sequence when working at extremely coarse sampling rates, whereas constant subsampling chose frames uniformly throughout the sequences. This tendency of activity-index subsampling was due to the large movements that occurred as the signer moved in and out of the rest position, producing the only $a_x(n)$ values that rose above $\alpha$. If this artifactual performance is discounted—that is, if the beginning and ending rest positions are removed from consideration—the estimate of overall activity-index superiority to constant subsampling increases to 15%.

The form of activity-index performance varies with stimulus presentation. For static presentation of full gray-scale images (Figure 6b), activity-index performance rises above that of constant subsampled images as the total number of frames decreases; the estimated difference in performance rises by nearly 20% when 6.75 fps are displayed. Activity-index performance, however, falls off sharply with fewer and greater numbers of frames per second. Interestingly, for activity-index sequences, FGS-S has a performance maximum at about 7 fps; for constant subsampled sequences, performance with FGS-S improves monotonically with frames per second. In contrast, static presentation of binary images (Figure 6d) produces flat and nearly equal performance for both subsampling schemes, reflected in the nonsignificant transformation factor in the analysis.

## Discussion

### Structure of Events

Although the experiment described here is not a direct test of the validity of Newtson's (1973) definition of breakpoints, there is certainly a close relation between their action-unit boundaries and our basis for dynamic sequence segmentation. Insofar as such a comparison may be made, the results reported here generally confirm findings of Newtson and his collaborators with regard to the perceptual salience of boundaries and their ability to convey critical event information. Indeed, in one condition, ASL sequences that were constructed via the activity index were as intelligible as the original sequence from which the frames were taken, despite a four-fold reduction in the number of frames. This result is similar to an intelligibility-rating result of Newtson and Engquist (1976) and yet, because of the objective nature of ASL intelligibility, is not open to the questions that follow the subjective rating paradigm used by Newtson and Engquist.

### Direct Perception of Events

A long-standing argument in theories of event perception revolves around the issue of whether events are directly per-

ceived, originating in Asch's (1952) theory that action-defining gestalten appear in the behavior sequence, or whether event perception is more of an interpretive, cognitive process. If action is directly perceived, it must be the case that the cues that give rise to the percept exist in the surface structure of the behavior sequence; that is, a necessary condition for the direct perception of events is that the basis for event structure must appear in the stimulus itself. This is the explicit assumption behind the behavioral segmentation method of Newtson (1973), and it is well supported by the many subsequent experiments by Newtson and his collaborators. If complete event information did not exist in the surface structure of the sequences used in the present experiment, it would have been extremely difficult, if not impossible, to segment our ASL movies into intelligible sequences. At the very least, some higher level, interpretive driver would have been necessary in order to produce compressed images that were more intelligible than those produced by constant subsampling. However, it is clear from the results of our experiment that the necessary event information does reside in the surface structure of sequences.

Even if it is conceded that events are directly perceived and that critical event information is carried by event boundaries, we would expect static presentation of behavior sequences to require a more interpretive process than does dynamic presentation of the same sequences. That is, events are usually not directly perceived with static presentation (Newtson & Engquist, 1976). Indeed, for dynamic presentation, the change in surface structure from one moment to the next is immediate, whereas for static presentations, the change must be inferred from an analysis of frames. That these are fundamentally different processes is reflected in the demonstration of left-hemispheric advantage for statically presented signs and the absence of lateral asymmetry for dynamic signs (Poizner, Battison, & Harlan, 1979). Because the inference process would certainly introduce an additional source of error, it seems likely that static images would be less efficient at transmitting the desired information. Accordingly, we note the generally lower intelligibility scores for static images in the present experiment.

The current experiment supports the notion that the basis for event structure appears in the stimulus itself. We found that the ASL events isolated by a simple image-based computation seem to agree with subjective impressions of event structure. This does not preclude the possibility that other sources of information, including higher reasoning, can act to modify the interpretation of an event. Nonetheless, our findings bode well for efforts in artificial intelligence directed toward machine interpretation of actions.

### ASL Primitives

A central component in the traditional study of ASL is the use of ASL primitives. Stokoe (1974) developed a set of primitives to describe signs that are composed of a limited set of movements, hand shapes, and locations of articulation. These components are meaningless when taken individually; when combined according to rule-governed constraints, they form the lexical basis of ASL. This is entirely analogous to

the function of phonemes in spoken language. A fourth ASL dimension, hand orientation, has subsequently been added to the list of primitives (Battison, 1974).

A somewhat remarkable result is the high intelligibility of the ASL sequences at extremely coarse sampling rates. Even in the most degraded condition, subjects correctly interpreted nearly a third of the signs. An explanation of these findings may stem from the relative importance of the four ASL primitives. Consider that a single frame taken from the middle of a sequence will likely convey information about three of the four primitives—hand orientation, hand shape, and location of articulation. Only motion is lost, or at the very least, severely degraded. The fact that subjects do so well with this limited amount of information reflects the degree to which nonmotion factors play a critical role in ASL intelligibility. Indeed, several studies have shown that the four primitives are not equally perceptible, nor are they equally important for intelligible ASL (Klima & Bellugi, 1979; Tartter & Fischer, 1982).

## Image Sequence Compression

*Activity index: Dynamic sequences.* It is apparent from the data and from subjective reports that for low frame-rate conditions, ASL sequences that have been subsampled with the activity index are more intelligible than those subsampled by a constant factor. In the full gray-scale dynamic condition (Figure 6a), activity-index sequences were correctly identified 80% of the time at slightly less than 11 new fps. When constant subsampling was used, the same performance level was not achieved until an estimated 20 to 25 new fps were displayed. At this criterion of performance, the number of to-be-transmitted frames was reduced by a factor of 1.8 to 2.25, roughly a twofold improvement over constant subsampling.

What are the implications of our findings? An 8-bit sequence with frames of 96 × 64 pixels shown at 30 fps requires 1.47 Mbits/s for full bandwidth transmission, more than 300 times the nominal capacity of the public switched telephone network. The large bulk of compression needed to transmit ASL sequences can certainly come from spatial compression and efficient data-encoding schemes, as demonstrated by Sperling et al. (1985). Nonetheless, sharing the effects of compression among spatial and temporal domains reduces the reliance on spatial compression, thereby reducing the amount of spatial information loss. Furthermore, it is easy to conceive of environments in which it is desirable for dynamic information to be transmitted, or encoded, as efficiently as possible. Intelligent temporal subsampling would have to be included in any such scheme.

The degree to which spatial and temporal compression may be joined depends on the degree of interaction between the two domains. Although activity-index subsampling yields sequences that are more intelligible than constant subsampling for the binary images used in the present study, the overall level of intelligibility, in relation to full gray-scale sequences, was reduced (although this particular comparison is across different groups of subjects). This interaction may suggest that there is limited promise in combining the two forms of compression. Indeed, Sperling et al. (1985) found that extreme

spatial compression yielded frames that were temporally decorrelated, so that additional temporal compression was ineffective.

It may be, however, that the particular form of spatial compression used in the present study undermined the success of activity subsampling. Our binary images were constructed by painting 10% of the pixels on the dark side of edges black on a white background, and the selection of pixels to darken might vary with slight changes in the signer's position. The physical representation of the signer within the sequence (i.e., the contours) emerged as a result of the juxtaposition of the black pixels averaged over several frames. Accordingly, a single frame taken from the middle of the sequence may represent the form of a human only very poorly; motion is necessary for the true physical structure of the signer to emerge. By disrupting the temporal characteristics of the sequences, we induced a breakdown of the spatial structure of the signer herself. Naturally, with the loss of spatial structure, intelligibility suffered. A better test of the temporal and spatial interaction would be to use a spatial compression scheme that preserves the structure within a frame without relying on motion cues and spatial averaging that occur between frames.

*Rotational and circular motions.* It was noted in the introduction that signs with rotational or circular motions present a unique problem to the sort of temporal segmentation conducted by the activity index: There are changes in the direction of the moving component (or components) without a corresponding change in velocity or acceleration. Depending on the criteria used to define a rotational or circular motion, between 6% and 15% of the signs used in this experiment could be so classified. By using the stricter criteria for inclusion, activity-index performance was compared with constant subsampling performance within a group of five signs with rotational or circular motions. There was no statistical difference between the two subsampling schemes within this group of signs. As expected, activity-index subsampling presented no advantage over constant subsampling for this group. In addition, intelligibility for the group of five rotational/circular signs was compared with intelligibility for the entire stimulus set. Although the difference was not significant, almost certainly a result of the small number of samples, intelligibility for the rotational/circular signs was, as a group, slightly lower than that of the complete set. Again, this is consistent with our expectations.

Despite the shortcomings of activity-index subsampling that appear when confronted with circular or rotational signs, the effectiveness of this technique is not likely to be greatly affected in any environment that more closely resembles the real world. In a continuous stream of signing, contextual constraints of the conversation will increase the overall intelligibility of the individual signs. Although there is a ceiling effect on many easily interpreted signs, these other, more difficult signs will be made more intelligible. Furthermore, we note that these signs usually represent a fairly small percentage of the total number of available signs.

*Application, real-time computation.* Sperling et al. (1985) demonstrated that telephone transmission of intelligible ASL was feasible; the experiments presented here indicate how this

minimal transmission can be improved by activity subsampling. In order for such a system to be useful to the signing public, the necessary hardware must be relatively affordable, easy to use, widely available, and the processing must be carried out in real time. In the present study, all computations were performed in software and required considerable computing power and time. However, the computations (the accumulation of frame-by-frame differences) were deliberately chosen to be of the kind that are easily embodied in parallel microprocessors. Indeed, we do not see any purely technical obstacles to producing video telecommunications devices that can transmit intelligible ASL over the ordinary switched communications network. Such facilities would have enormous practical significance for the signing deaf and hearing impaired, reducing their isolation from each other and, one hopes, from the hearing community at large.

## Static Presentation

*Optimal number of frames.* Two interesting findings emerge from the static presentation conditions. First, it is encouraging to note that even in the most difficult condition, there was still a 30% chance of correctly identifying the presented sign. This attests to the robustness of the ASL signal. Second, as noted in the Results section, performance for static presentation of full gray-scale signs declines when there are more than 6 fps in the activity sampling condition. Why?

Subjects reported difficulty with the task of scanning through a page of "printed" ASL frames, although they improved with practice. The most common complaint was that there were "too many frames to see what was going on." If it were simply the case that there was an optimum number of frames for each sign, then we would have expected to see evidence of this in both subsampling conditions. Yet, this pattern emerged only for activity-index subsampling. The difference is that although activity-index subsampling chooses critical frames, when the frame repetition factor *m* is increased in constant subsampling, critical frames are just as likely to be discarded as any other frames. For constant frame-rate sampling, the improved performance that would have resulted at the optimal frame rate is compensated by the loss of critical information. It is not just that there is an optimal number of frames, but that there are optimal frames, and that activity index subsampling is one method of discovering optimal subsets of frames.

*Automatically generated ASL text.* The ability of subjects to "read" static signs and our ability to use digital image technology to produce static text raise an intriguing possibility: messages or even books composed entirely of signed sequences. The automatic production of such static text offers ASL signers an opportunity for veridical representation of ASL conversations that is understandable directly without mechanical aids, such as VCRs. Direct quotes, jokes, announcements, and the like can be communicated with individual expression and intonation. It remains to be determined whether signers could, with practice, become sufficiently proficient at reading signed text to make these possibilities practicalities.

## References

Abramatic, J. F., Letellier, P. H., & Nadler, M. (1982). A narrow-band video communication system for the transmission of sign language over ordinary telephone lines. In T. S. Huang (Ed.), *Image sequence processing and dynamic scene analysis* (pp. 314–316). New York: Springer-Verlag.

Asch, S. (1952). *Social psychology.* Englewood Cliffs, NJ: Prentice-Hall.

Battison, R. (1974). Phonological deletion in American Sign Language. *Sign Language Studies, 5,* 1–9.

Bellugi, U., & Fischer, S. (1972). A comparison of sign language and spoken language. *Cognition, 1,* 173–200.

Cutting, J. E. (1978). A program to generate synthetic walkers as dynamic point-light displays. *Behavior Research Methods and Instrumentation, 10,* 91–94.

Ebbesen, E. B. (1980). The development of behavior perception. In R. Hastie, T. Ostrom, E. B. Ebbesen, R. S. Wyer, D. L. Hamilton, & D. E. Carlston (Eds.), *Person memory* (pp. 46–69). Hillsdale, NJ: Erlbaum.

Green, K. (1984). Sign boundaries in American Sign Language. *Sign Language Studies, 42,* 65–91.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Humphries, T., Padden, C., & O'Rourke, T. J. (1980). *A basic course in American Sign Language.* Silver Springs, MD: T. J. Publishers.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201–211.

Klima, E., & Bellugi, U. (1979). *The signs of language.* Cambridge, MA: Harvard University Press.

Landy, M. S., Cohen, Y., & Sperling, G. (1984a). HIPS: A Unix-based image processing system. *Computer Vision, Graphics, and Image Processing, 25,* 331–347.

Landy, M. S., Cohen, Y., & Sperling, G. (1984b). HIPS: Image processing under Unix, software and applications. *Behavior Research Methods, Instrumentation, and Computers, 16,* 199–216.

Levitan, E. L. (1960). *Animation art in the commercial film industry.* New York: Reinhold.

Markus, H., & Zajonc, R. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol. 1, pp. 137–230). New York: Random House.

Marr, D., & Vaina, L. (1980). *Representation and recognition of the movements of shapes* (Tech. Rep. No. 597). Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28,* 28–38.

Newtson, D., & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology, 12,* 436–450.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35,* 847–862.

Newtson, D., Hairfield, J., Bloomingdale, J., & Cutino, S. (1987). The structure of action and interaction. *Social Cognition, 5,* 191–237.

Newtson, D., & Rindner, R. (1979). Variation in behavior perception and ability attribution. *Journal of Personality and Social Psychology, 37,* 1847–1858.

Parish, D. H., & Sperling, G. (1987). Object spatial frequency, not retinal spatial frequency, determines identification efficiency. *Investigative Ophthalmology and Visual Science, ARVO Supplement, 28,* 359.

Pearson, D. E. (1981). Visual communication systems for the deaf. *IEEE Transactions on Communications, 29,* 1986–1992.

Poizner, H., Battison, R., & Harlan, L. (1979). Cerebral asymmetry for American Sign Language: The effects of moving stimuli. *Brain and Language, 7,* 351–362.

Poizner, H., Bellugi, U., & Lutes-Driscoll, V. (1981). Perception of American Sign Language in dynamic point-light displays. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 430–440.

Riekehof, L. L. (1980). *The joy of signing.* Springfield, MO: Gospel Publishing House.

Rindner, R. (1982). *Different units of information in the perception of behavior.* Unpublished doctoral dissertation, University of Virginia, Charlottesville.

Rubin, J. M., & Richards, W. A. (1985). *Boundaries of visual motion* (Tech. Memo No. 835). Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Sperling, G. (1978). Future prospects in language and communications for the congenitally deaf. In L. Liben (Ed.), *Deaf children: Developmental perspectives* (pp. 103–114). Orlando, FL: Academic Press.

Sperling, G. (1980). Bandwidth requirements for video transmission of American Sign Language and finger spelling. *Science, 210,* 797–799.

Sperling, G. (1981). Video transmission of American Sign Language and finger spelling: Present and projected bandwidth requirements. *IEEE Transactions on Communications, COM-29,* 1993–2002.

Sperling, G., Landy, M. S., Cohen, Y., & Pavel, M. (1985). Intelligible encoding of ASL image sequences at extremely low information rates. *Computer Vision, Graphics, and Image Processing, 31,* 335–391.

Stokoe, W. C., Jr. (1974). Classification and description of sign languages. In T. A. Sebeok (Ed.), *Current trends in linguistics* (pp. 345–371). The Hague, The Netherlands: Mouton.

Stokoe, W. C., Jr., Casterline, D., & Croneberg, C. G. (1976). *A dictionary of American Sign Language.* Silver Springs, MD: Linstok Press.

Tartter, V. C., & Fischer, S. D. (1982). Perceptual confusions in ASL under normal and reduced (point-light display) conditions. *Perception & Psychophysics, 32,* 327–334.

Thibadeau, R. (1983). *Artificial perception of actions.* Unpublished manuscript.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.