Sperling, G., Pavel, M., Cohen, Y., Landy, M. S., & Schwartz, B. J. (1983). Image processing in perception and cognition. In O. J. Braddick & A. C. Sleigh (Eds.), *Proceedings of Rank Prize funds International Symposium at The Royal Society of London, 1982. Springer Series in Information Sciences: Vol. 11. Physical and Biological Processing of Images* (pp. 359-378). Berlin: Springer-Verlag.

Image Processing in Perception and Cognition

G. Sperling, M. Pavel, Y. Cohen, M.S. Landy, and B.J. Schwartz

New York University, New York, N.Y. 10003, USA

1. Overview

In this paper we briefly survey theories and ideas about image processing, with some illustrative examples, taken mostly from the Human Information Processing Laboratory at N.Y.U. First we develop the concepts of *multiple stable states* and *path dependence* in a basic visual-motor task (vergence of the eyes) and show how these can be encompassed in *potential theory*. We then consider two examples of human information extraction from complex, dynamic, visual displays: (1) the extraction of the shape and the motion of 3D wire objects from 2D images, and (2) the extraction of meaning from displays of deaf signers communicating in American Sign Language (ASL).

In extracting 3D shape, the human perceptual system uses *fallible heuristics* which can lead to obvious perceptual errors (i.e., powerful perceptual illusions). Further, in ambiguous stimuli, the induced perceptual state depends strongly on immediately preceding stimuli *(path dependence)*. The theory of the "logic of perception" proposed to account for these observations is that information about a stimulus--derived concurrently by many different processes--is combined in a nonlinear competition/cooperation network to achieve a perceptual decision.

In cognitive information processing tasks, such as extracting meaning from continuous ASL communication, there are two levels of processing: first, parsing the continuous sequence of images into a sequence of signs (analogous to parsing spoken language into words or cursive script into letters) and second, extracting meaning from the sign sequence. The parsing problem is characterized by the extremely restrictive structure of the information received. For example, the number of different signs in ASL is only on the order of thousands. Therefore, experienced signers can extract the reference form of a sign from very reduced images and from very different contexts (even though performance of the same sign may vary considerably with context). A practical consequence is that we have been able to produce legible ASL images of very low bandwidth (on the order of 10 kHz) by various image-coding schemes: spatial low-pass filtering, binary intensity encoding, and cartooning by edge detection or zero-crossing algorithms. *Feature templates* and *intentions* are the operative concepts for parsing; *schemas* are the most hopeful approach to the problem of extracting meaning.

2. Image Processing in Perception

2.1. Multiple stable states and path dependence

2.1.1. Multistability of vergence. We consider first a classical demonstration of HELMHOLTZ [17] that may seem specialized but is prototypical of human perception. A subject views a binocular stereogram, that is, a stimulus that produces images that are identical--or nearly identical--at corresponding points in each eye (Fig. 1). To normal subjects, such a stimulus appears as a single object, that is,

the images in the two eyes are said to be perceptually fused. Suppose, at this point, the lines of sight of the two eyes are parallel. Now, let the left and right half-images of the stereogram be slowly moved apart in physical space so that, in order to maintain fusion, the lines of sight of the two eyes must diverge. Helmholtz found that he was able to cause his eyes to diverge by eight degrees before fusion was lost and his eyes returned to their resting position. To restore fusion, Helmholtz had to bring the stereo half-images much closer together than the point at which fusion was lost.

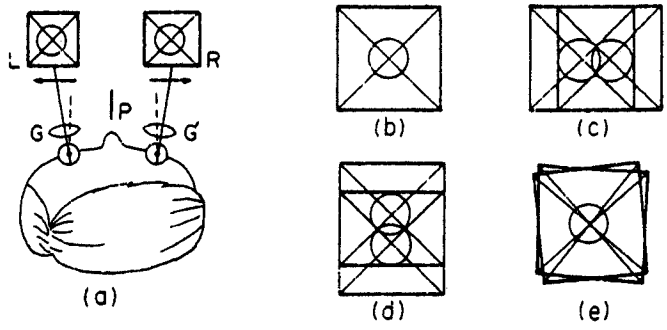

Fig. 1. (a) A stereogram showing the left *L* and right *R* half-images viewed by the left and right eyes, respectively. The lenses *G*, *G'* enable the observer to focus at the short viewing distance. The partition *P* isolates the views of the two eyes. The arrows indicate the direction in which the stimuli are moved in Helmholtz's horizontal vergence demonstration. The broken lines indicate parallel lines of sight; the solid lines indicate the actual lines of sight when the observer is able to verge accurately at the angle of divergence shown. (b) Representation of the fused perception when the eyes are correctly verged. The perception after vergence fails for (c) horizontal vergence, (d) vertical vergence, and (e) torsional vergence. (After Sperling [42])

Some people can control horizontal vergence voluntarily, and this will complicate any demonstration. However, analogous results to those described above are obtained with vertical and with torsional displacements. That is, when the stereo half-fields are displaced vertically the eyes can be forced to diverge vertically by up to about eight degrees, and with torsional rotation of the fields, the eyes can be forced to rotate torsionally around their axes (Figs. 1d,e). To the best of our knowledge, no one has yet succeeded in *voluntarily* making vertical or torsional eye movements.

2.1.2. Multiple stable states, path dependence, hysteresis. These demonstrations from Helmholtz illustrate path dependence for responses to a certain range of stimuli. When a stereogram requires an ocular divergence of, say, one degree, then divergence always occurs and fusion is always achieved, independent of the recent history of stimuli. With a stereogram requiring a divergence of twelve degrees, fusion is never achieved. With a stereogram requiring a vergence of eight degrees, fusion will occur if the immediately preceding stimulus requires a divergence seven degrees and the eyes were fused on it but not if the preceding stimulus divergence was one or twelve degrees. There are two stable states, fused or unfused, in response to precisely the same stereogram. We say there

are *multiple stable states* and there is *path dependence*. Insofar as we restrict ourselves to deterministic theories, multiple stable states and path-dependence are two sides of the same coin: different states can be reached only by different paths (sequences of previous stimuli) and, path-dependence implies that the same stimulus induces different states depending on the path taken. *Hysteresis* is the name given to path dependence in which a state tends to perseverate even after the inducing conditions have changed. This is the kind of path dependence demonstrated above.

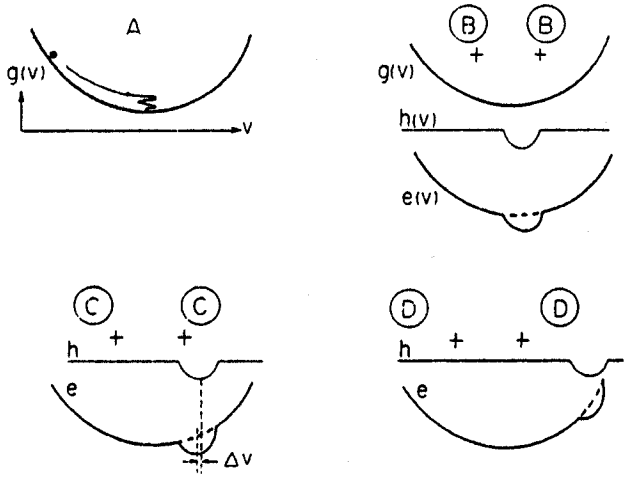

Fig. 2. Potential theory model. (A) Vergence displacement energy $g(v)$ as a function of vergence angle v. The eye's vergence position is represented by the projection onto the abscissa of a marble rolling on the surface. This surface governs the eyes' return to their neutral vergence position when they are somehow displaced from it; a typical path is indicated. (B) Two halves of a stereogram (B,B), fixation points are indicated by +. Vergence displacement energy $g(v)$ adds to image-disparity energy $h(v)$ to produce (net) vergence energy $e(v)$, which has a single minimum slightly displaced in the direction of divergence. (C) Two halves of a stereogram requiring a greater divergence than (B) to achieve fusion; $g(v)$ is same as in (B) and therefore omitted; the corresponding $h(v)$ is shifted to the right. The e(v) surface has two minima. The arrows indicate "vergence disparity" Δv: the displacement of the minimum of $e(v)$ [which determines the actual vergence position] away from the minimum of $h(v)$ [which minimizes *L/R* image disparity]. (D) A stereogram which is at the limit of vergence fusion. The $e(v)$ surface has only one minimum, which corresponds to the neutral vergence position. (After Sperling [38])

2.2. Potential theory interpretation of path dependence.[2]

Let the vergence position, v, of the eyes as a function of time, t, be given by v(t). Let the composite of all forces tending to alter vergence position be dg/dv, and let

$$dg/dv = -k_1 v'' - k_2 v' \qquad (1)$$

where ' denotes differentiation with respect to t. Equation 1 is an elementary, linear differential equation in which the coefficient k_1 can be thought of as describing the mass of a moving body and k_2, the friction of the medium in which it moves. A simple, concrete realization of (1) would be a salad bowl filled with oil in which a marble rolls under the influence of gravity. The shape of the bowl is described by g(v) (Fig. 2a). The vergence position of the eyes is represented by the horizontal coordinate of the marble. In electrical potential theory, g represents field potential, dg/dv represents force; here, g is called *vergence displacement energy.*

For bowl-shaped g(v) and nonzero friction ($k_2 > 0$), it is evident that, whatever the starting position and velocity of the marble, it must eventually come to rest at the bottom of the bowl. (We assume the conditions are such that the marble remains in the bowl.) This salad bowl model represents the internal control of vergence movements of the eyes, e.g., their movement in the absence of any external stimulus. For example, suppose the eyes are verged on a particular stereogram and suddenly the lights are turned off. This is represented by positioning the marble at the side of the bowl and allowing it to roll freely to its resting position (Fig. 2a).

External factors in the control of vergence are represented by image disparity energy, h(v). h(v) is expressed in terms of the squared differences between the illuminance distributions on the left and right *retinas*, which in turn, are expressed in terms of the luminances of the *stimuli* to the left and right eyes, $I_L(x,y)$ and $I_R(x,y)$, and the vergence position v of the eyes:

$$h(v) = -\iint_{x,y} \left[I_L\left(x - \frac{v}{2}, y\right) - I_R\left(x + \frac{v}{2}, y\right) \right]^2 dx\ dy. \qquad (2)$$

One should think of *h* as the error in image registration--the square of the difference between the two eyes' images. The critical point is the assumption that h(v), or some other error function like it, can be computed not only for the eyes' present vergence position but for other values of vergence. It is easy to imagine several ways in which neurons might compute h(v) in the neighborhood of the current vergence, v [38]. Insofar as h(v) can be computed in a neighborhood of v by the visual system, it would know in which direction to move the eyes to reduce the registration error h(v), i.e., to increase the correspondence of the two left and right retinal images. Figure 2b illustrates a typical h(v).

In the potential-theory model, vergence is controlled by e(v) = g(v) + h(v), that is, by the sum of internal and external factors, i.e., the marble rolls on the e(v) surface. How this works is illustrated for three cases. When a stereogram requires only a small amount of divergence for fusion, there is a single minimum in e(v) and consequently only one stable state: fusion of the object. When the stereogram requires a greater amount of divergence, there are two stable states: resting position and fused on the stereogram. From e(v) in Fig. 2c it is clear how, by moving the minimum of h gradually from the position represented in 2b to that in 2c, the marble will stay in the minimum corresponding to the fused state. Figure 2d illustrates the case of too great a disparity for fusion, the perturbation in e(v) caused by h(v) is too shallow to hold the marble.

Figure 2 illustrates the potential theory representation of the multiple stable states of vergence. There is also a catastrophe theory representation (SPERLING [42]) which is more succinct but omits any suggestion of the dynamics of the system. The vergence system is an especially attractive one in which to investigate dynamics because all the intermediate states of the system as well as the stable states are observable.

2.3. Neural model for multiple stable states

Like vergence, higher level perceptual processes in humans (and presumably, in all biological systems) also are characterized by path dependence and multiple stable states. The functional description of these processes is quite similar to the case of vergence, but the actual mechanism involves, we believe, a very specialized neural network for making decisions. In horizontal vergence, the eyes can be verged on only one vertical plane at any one time, and are verged on exactly one plane because of physical constraints.

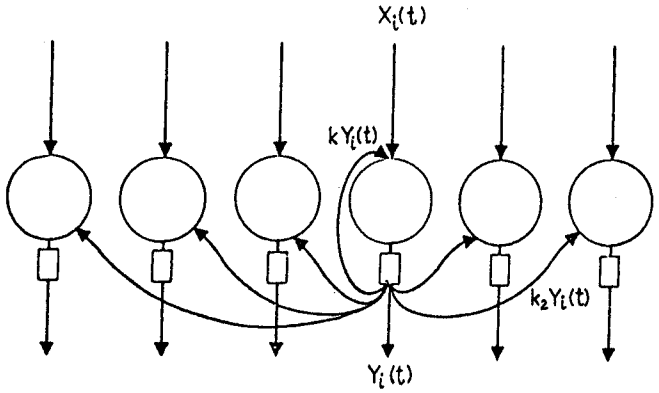

Fig. 3. A monoactive column of model neurons. Connections are shown for one neuron, i. It receives an external input $x_i(t)$ that sums with an output-produced feedback excitatory input $k_1 y_i(t)$. It sends shunting inhibitory signals $k_2 y_i(t)$ to all other cells. The small rectangular box in the output path represents a threshold: when its input is y, $y > 0$, its output is $\max(y-\theta, 0)$. (After Sperling [42])

An analogous neural network is the "monoactive" network of n neurons, only one neuron can be "active" at one time and exactly one neuron is active at every time. A proposed wiring diagram for a monoactive net is shown in Figure 3. Neurons are labeled i, inputs are $x_i(t)$, outputs are $y_i(t)$, each neuron feeds back its output onto itself as an excitatory input and every neuron sends its output to all the other neurons in the net as a strong inhibitory input. Simple inhibitory networks are not monoactive. Three properties make this inhibitory network monoactive: (1) the output range of neurons is bounded; (2) neurons have a threshold, ϵ, which their net input must exceed in order for there to be an output; (3) each neuron has positive self-feedback. The first property holds for all neurons, the second for nearly all, and the third is unusual. A network with these properties was proposed by SPERLING [38], and independently by GROSSBERG [15, 16] who developed powerful mathematical analyses of such systems. A monoactive network is necessary where decisions need to be made between alternatives that cannot or should not be combined. For example, when there are several widely-spaced objects competing for attention in the visual field, the eyes point at each of the objects in turn, not at the mean position, which may be blank. When an animal is hungry, thirsty, and sleepy, and must satisfy these drives at different locations, it goes to each location in turn, not to an in between location.

2.4. Cooperation/Competition

We consider here three processes that occur in the human cyclopean image, that is, the perceived image of the world produced by the two eyes acting together. The thrust of these examples is that various decisions are made about each point in the cyclopean field. Each decision mechanism can be represented as a monoactive column of cells centered over x, y. Although the interaction within a column is entirely competitive, neighboring columns are assumed to interact cooperatively. That is, a decision reached in one column favorably influences the analogous outcome in its neighbors (i.e., the active neuron provides an excitatory input to neurons on the same level in neighboring columns). Between-column interactions are local, but effects can propagate widely.

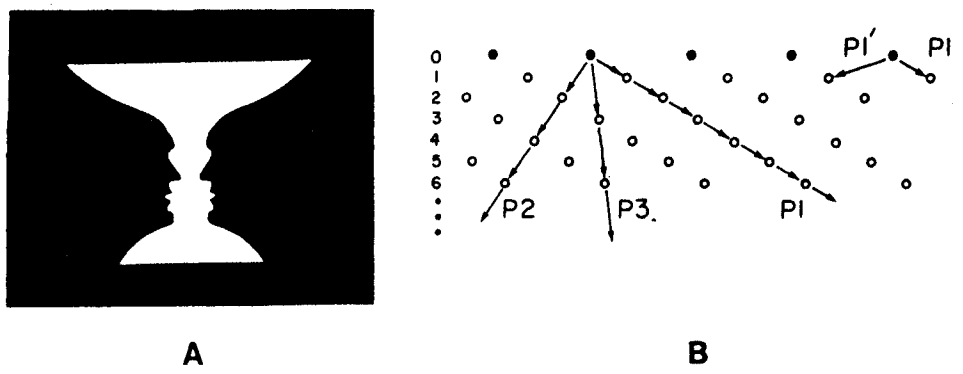

Fig. 4. Two examples of ambiguous displays. (A) Static figure/ground ambiguity, Rubin's faces/vase. (B) Ambiguous motion display. Dot row zero is flashed first, followed by row 1, 2, and so on. Motion can be perceived along any of the paths P1, P2, P3; which of these paths dominates in this configuration depends on the time t and the distance d between successive points along the path. Although P1' is a potential pathway, it is not perceived because it has the same t as P1 but a larger d, and is therefore (theoretically and empirically) always weaker. (Based on Burt & Sperling [4].)

2.4.1. **Stereoscopic Depth.** At any point x,y in our cyclopean image we perceive one and only one value of depth. That is, we cannot see two fine details at different depths at precisely the same point in the cyclopean field. Even in looking at a landscape through a lace curtain, each point will be seen to belong either to the landscape or the curtain. In stereograms containing large regions of ambiguous depth, the perceptual resolution of these regions can be determined by a few points of unambiguous depth.

2.4.2. **Binocular rivalry.** When conflicting features (e.g., lines at different orientations) are presented to corresponding points of the two eyes, binocular rivalry is usually observed--that is, a location x, y of the cyclopean field will contain the feature of one eye or the other, but not both. Large ambiguous (rivalrous) regions are easily influenced by small unambiguous regions, and tend to resolve as wholes.

2.4.3. Figure–ground. In the classical ambiguous figures (Fig. 4a), every point in the cyclopean image is labelled as "figure" or "ground". Any region that is unambiguously figure tends to expand to its limits, as does any region labelled ground. In the classical ambiguous figures, there are two equally valid, symmetrically opposite labelling schemes. The figure-ground labelling process probably is a manifestation of a more general object labelling process that assigns each point of the visual field to an object or to the background.

Since SPERLING [38] (proposed cooperation/competition networks to account for phenomena of stereoscopic depth perception, they have been widely adopted by theorists in similar contexts (e.g., JULESZ [22], DEV [11], NELSON [30], MARR & POGGIO [28]).

2.5. Perceptual Selection: The Dominance Problem

The examples up to this point have dealt largely with the mechanism by which perceptual decisions are made and have not considered the nature or situation-specific content that enters into these decisions. To open this discussion, let us consider a three-dimensional wire cube rotating about a vertical axis. When viewed monocularly, such a cube is ambiguous: it can be perceived either *veridically* (correctly) or *reversed*, (i.e., the front face appears in the rear, the rear face in front, and the apparent direction of rotation is reversed). To avoid the influence of incidental cues (shadows, texture and imperfections in the wires, the slight increase in angular subtense of the wire edges as they approach the observer, etc.) it is best to represent the cube dynamically on a CRT screen. Ideally, this CRT display would produce the same monocular retinal image as a real wire cube except that the CRT's visual properties are more flexibly controlled than those of real wire cubes. Thus, it is not surprising that a well-constructed 2D CRT display can look virtually as solid, as 3D, and as convincing as a real wire cube. When our subjects' visual perception of a CRT display spontaneously "flips" from the veridical to reversed state (i.e., the direction of motion reverses), it is so convincing that they mistake the perceptual state change for an objective change in direction.

There are two critical questions relating to the perceptual reconstruction of 3D wire figures from their 2D projections. (1) How are the various alternative perceptual states computed? (2) What determines the relative dominance of the perceptual alternatives? We consider the question of dominance first.

2.5.1. Control of dominance by proximity luminance covariance (PLC). In the case of parallel projection of a rotating wire object (Fig. 5a), the veridical and the reversed figures are mirror images of each other and appear to rotate in opposite directions. There is no physical basis for discriminating between the two alternatives. In experiments, subjects report the two alternative states with equal frequency. Suppose we introduce a positive PLC into the display, that is, we make edges that represent portions of the 3D cube close to the observer brighter than those edges that represent distant portions (Figs. 5c, g). Now we find (SCHWARTZ & SPERLING [37]) that 96% of the time the observers report the perceptual alternative with brighter edges forward; only 4 percent of the time are bright edges seen in the rear. In the case of parallel projection, making *distant* parts of a cube brighter (a negative PLC) does not produce a different result because, by symmetry, the stimulus is objectively identical to that obtained by making near parts brighter. (Only the direction of rotation is reversed.)

To test the strength of PLC as a cue to shape, we can pit it against other factors, such as rigidity. Consider a polar projection of a cube, such as that

obtained by viewing a cube from close up, say, from a vantage point whose distance to the center of the cube is only 2.25 times the width of a side. This 2D perspective view is illustrated in Fig. 5e. When such a wire cube is rotated, the shape ambiguity is removed. There is one, and only one, interpretation as a rigid object. Nevertheless, as with parallel projection, the cube can easily be perceived in two states: veridical and reversed. With polar projection, only the veridical state is perceived as rigid; the reversed state is perceived as a nonrigid, rubbery object whose sides distort enormously, with nonuniform expansions and contractions throughout the rotation. One might suppose that subjects have a strong tendency to perceive the object in the rigid mode, but empirically, Schwartz and Sperling found this not to be the case.

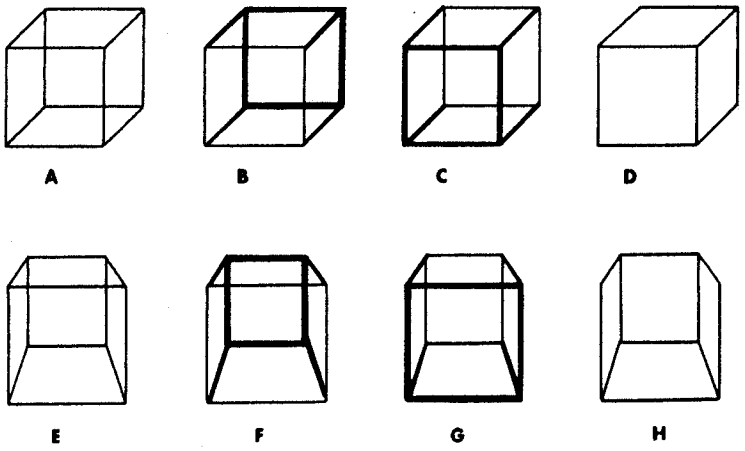

Fig. 5. Frames from rotating Necker cubes. (A) Necker cube. (B) Positive *proximity luminance correlation* (PLC). (C) Negative PLC. (D) Hidden lines (occlusion). (E) Perspective cube. (F) Positive PLC. (G) Negative PLC. (H) Hidden lines (occlusion). (Based on Schwartz & Sperling [37])

There is a modest tendency, for the rigid mode dominates slightly over the nonrigid mode, but this seldom exceeds about 70:30 and depends irregularly on such factors as polar projection distance and subject, with some subjects perceiving the nonrigid mode more often. It is not surprising then, that PLC overwhelms rigid interpretation (in linear perspective) as a determiner of dominance. With a positive PLC that favors the veridical mode (Fig. 5g), the rigid percept occurs on 94% of the presentations. With a negative PLC that favors the reversed mode (Fig. 5f), the *nonrigid* percept occurs on 92% of the presentations. The significance of these demonstrations of the power of PLC is that PLC is a fallible cue to depth that nevertheless consistently overrides unequivocal geometrical cues. It might be well to design machines to use more reliable cues to depth than humans do.

2.5.2. **Other selection cues.** Proximity luminance covariation is the most powerful cue we have discovered in wire cubes. By reducing the PLC (so that the brightness ratio of near points to far approaches 1.0), the strength of the PLC

cue can be made arbitrarily small, and equated to any other cue, such as linear perspective that may be pitted against it. This balance method, and the corresponding analytic techniques of conjoint measurement [24] [12] can be used to accurately scale the strength of cues that determine the choice of perceptual states, and ultimately to construct an additive scale for the strength of any combination of factors. For example, *occlusion* (the removal of lines that would be hidden if the sides of the cube were opaque) is a weak cue to the perception of reversed states because, in polar projections, very little is actually occluded (Fig. 5h), but a strong cue to rigid perceptions because much is removed (not shown, cf. Fig. 5d). On the other hand, by painting opaque blotches on the side of a figure, occlusion can be made overwhelming (BRAUNSTEIN [3]). Linear perspective, as we noted above, is a relatively weak cue. The effect of a preview of a static field is quite strong. Context (adjacent or surrounding cubes) is not particularly effective.[4]

2.5.3. Weighing the evidence. The theory that we derive from experiments of this sort, is that there are a small number of discrete perceptual states, and that the evidence in favor of the states is simply weighed linearly, with the most favored state dominating in proportion to the weight advantage of its evidence over that of the strongest contender(s). Two comprehensive recent studies take a similar approach.

(1) The late Frank RESTLE [34] introduced a mathematical strength theory to the study of motion in his seminal analysis of JOHANSSON'S [19] fascinating motion stimuli. In a typical stimulus configuration, the movement of a small number of dots along interrelated linear and elliptical paths yields a rich variety of possible percepts (2D and 3D interpretations). Restle's strength function counts the number of parameters that are necessary, in his system, to describe the trajectories of dots in each of the candidate perceptual modes. The smaller the number of parameters--the simpler the description of the perceptual state--the greater its strength.

(2) BURT and SPERLING [4] studied ambiguous motion in stimuli composed of rows of dots (Fig. 5b). They found that the time t and distance d between dots determined the path, among several, along which motion was perceived. They derived a concise additive representation for the strength S of any candidate path of the form

$$S = te^{-t/\lambda}/d\,, \tag{3}$$

where λ was about 20 msec. (This representation becomes additive by taking logarithms of both sides of (3)). By functional analysis, Burt and Sperling were able to prove that (3) was a unique description of their data.

In the past, investigators have not looked for additive representations for the weight of sources of evidence in favor of perceptual hypotheses. On the other hand, linear estimates are the simplest, and in many instances the most powerful statistical estimates of the significance of evidence, and almost certainly among the easiest to implement in machines. The monoactive neural network suggests that biological implementation also may be simple. Indeed, we may be entering an epoch of convergent evolution of psychological theories, biological theories, and machine implementations.

2.6. Perception Selection: The Selection Problem

How does the perceptual system arrive at the candidate perceptual states for which it seeks evidence? In the case of rotating wire figures, there are three proposals to consider.

(1) ULLMAN [46] demonstrated that by identifying (the same) five points of a 3D object in three different 2D projections, the 3D configuration of the points could be computed algebraically. This is a global computation of depth in the sense that no bit of information in any single frame conveys any 3D information; the 3D emerges only when the whole computation, involving all the five points under consideration, is complete. So far, this algorithm has not been elaborated to deal with errors, even small errors, but it is constructive and demonstrably works well with error-free data.

(2) TODD [45] proposed a theory quite analogous to RESTLE'S [34] in which motion is analyzed in terms of projections of ellipses in 2D representing conic sections in 3D. This approach is promising but lacks generality.

(3) We propose that local cues provide inputs to an interactive competition/cooperation network whose state tends toward the most self-consistent percept. Examples of local cues are shrinking or expanding lines, changes in direction of movement, etc. in dynamic views, and vertices, converging lines, etc., in both static and dynamic views.[4]

Although in formal models, the selection and dominance problems can be treated separately, in biological systems, selection and dominance undoubtedly are two manifestations of the same computation.

3. Image Processing in Cognition

There are no hard and fast differences between perception and cognition: we use the term *cognition* to refer to complex situations in which learning, language, and meaningfulness play a greater role.

3.1. Bandwidth of American Sign Language (ASL)

Although we think of language communication as one of the more complex of cognitive tasks, there is an old English saying that belittles language relative to images; it translates into modern English roughly as

An image has the utility of 10^3 speech tokens. *(4)*

In the next sections, we investigate the extent to which (4) is indeed true.

In the U.S.A., the bandwidth of AM radio transmission is about 4.5 kHz. That is, radio transmits acoustic signals from about 0.3 to about 4.8 kHz. American television was designed for a bandwidth of over 4 MHz [31] so that we find Eq. (4) verified. Similarly, the bandwidth of a telephone line in the U.S.A. is approximately 3 kHz [47]. American Picturephone and British Viewphone, systems for face-to-face video telecommunication, were designed with bandwidths of over 1 MHz [9], [18]. Again, Eq. (4) is near the truth. Unfortunately, today the cost of a communication line is closely related to its bandwidth, so that video communication costs orders of magnitude more than speech communication.

The high cost of video communication is of little concern to hearing persons who normally use telephones for communication. It is a tragically insurmountable obstacle to congenitally deaf persons who habitually communicate in sign language; they can not use telephones and could not afford video telephones. In this regard, studies of proficient users of American Sign Language (ASL) show that they communicate information at about the same rate as in spoken language [1], so that the signing population could use a video telephone as effectively as the hearing population uses telephones. No one had ever measured the bandwidth actually required to communicate in ASL by means of raster encoded images, and when no one responded to a suggestion to do so [39], the first author set about to do it himself.

The ordinary television screen encompasses a nominal bandwidth of 4 MHz. By dividing the screen into a large number of rectangles, like a sheet of postage stamps, a different ASL conversation could be carried in each postage-stamp rectangle. By varying the size of the stamps, and calculating the fraction of total bandwidth allocated to each, SPERLING [40][41] was able to show that with a bandwidth of 21 kHz, ASL sentences and ASL word lists were communicated at 90% of the intelligibility of the control condition and finger spelling was at 70% of control. The most experienced deaf subjects achieved intelligibility scores of 40% to 50% at a bandwidth of 4.4 kHz. This is about what would be obtained with hearing subjects listening to voice communication over a voice communication channel with a bandwidth of 1.5 kHz [13]. Thus, while a television picture may use 1000 times the bandwidth of a telephone, ASL requires only a few times more bandwidth than voice communication. There are some obvious next questions. (1) Is it possible to use image processing methods to reduce the picture bandwidth so that ASL could be communicated over ordinary telephone lines, or, if not, what is the requirement for an ASL communication line? (2) How is it possible for experienced signers (or speakers) to use degraded signals so effectively?

3.2. Image processing of ASL

The intelligibility of degraded images indicates that a sequence of good quality ASL images contains a large number of redundancies. That is, some of the information present at a given point of the image is also present elsewhere. A human observer is capable of taking advantage of the redundancies and extracting the important information from a quite impaired image. Solution of the transmission problem requires a machine to take advantage of these redundancies to transmit an efficient code. There are three potential sources of redundancies:

(1) within-frame statistical and deterministic dependencies

(2) between-frame statistical dependencies

(3) source constraints.

These three cases overlap. The first two cases involve assumptions concerning only the images or train of images without any additional higher level knowledge. The last case involves assumptions regarding the physical, semantic and pragmatic constraints. That is, the signer is limited by her anatomical structure combined with the laws of physics, by the rules of the language, and by the message to be conveyed. For example, Fig. 6 shows a selection of frames from the sign "tomato". In "tomato" the index finger first touches the lips (indicating taste), then rotates out and imitates a slicing movement. There are large relatively homogeneous areas within all the frames of Fig. 6; some areas hardly change from frame to frame; when the head or body moves, it moves very slowly; details in the right hand are important for conveying meaning primarily when it is still or moving very slowly; details in the left hand are unimportant, and so on. These facts could and should be embodied in an efficient coding scheme.

3.2.1. Nondestructive Coding. The studies of the parameters of raster coding of ASL [32][40][41] did not enable us to answer the most interesting questions, and it was necessary to establish a digital image processing facility--a substantial technical undertaking (see Acknowledgements and LANDY, [25]; COHEN, [6][7]). Our basic digitized images of ASL were 96 pixels (picture elements) high by 64 pixels wide and quantized to seven significant bits of intensity (128 discriminable grey levels). There were a maximum of 30 full frames per second, with interlace. In this paper we concentrate on the preliminary results rather than the technical details. Our first step in the search for a more efficient image representation is

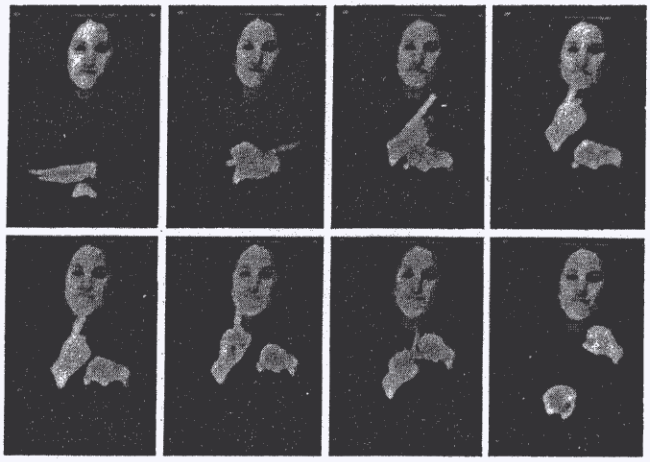

Fig. 6. Frames from a 1.7 sec (50 frame) production of the ASL sign "tomato". Beginning at upper left, the frame number in sequence is 3, 13, 17, 21, 28, 36, 45, 49. The frames are 64x96 pixels, with 7 bits of intensity resolution in the original display. The signer is wearing a dark sweater and viewed against a black background in order to facilitate discrimination of the hands. The white dots in the corners are fiducial marks placed just outside the corners of the picture area

nondestructive coding. Here, we represent the images by a code which is shorter for frequently occurring symbols or which combines a number of symbols into one.

3.2.1.1. *Huffman codes.* One such code, based on Huffman sequences, treats each pixel individually. Each value of a grey level is converted into a binary word whose length is inversely proportional to the relative frequency of occurrence of that level. Thus, the frequently occurring values of luminance are represented by only a few bits and the rare values by many bits. Significant reduction of information is achieved in those cases where the distribution (histogram) of the gray levels differs considerably from a uniform distribution. When a coding of this type is applied to the ASL images, the amount of data compression is relatively small. The theoretical limit of such coding schemes may be estimated by computing the entropy of individual pixels for the ASL images. For our sample of the ASL images, the entropy indicated maximum possible compression to be 38% of the original value.

3.2.1.2. *Run length codes.* Another nondestructive code which appeared promising is the run-length code. Instead of representing each pixel, this code represents any sequence of pixels of the same value of gray level by the level and the length of the sequence. Consequently, the run-length code is particularly powerful when the images contain considerable patches of the same luminance level. In the case of the ASL images with full grey scale, the savings were nominal; i.e., 70% of bits were required.

3.2.1.3. *Hierarchical codes.* There are many other, more sophisticated, methods for information compression. An attempt was made to apply some of them, such

as hierarchical coding [23], but the amount of compression achieved was small in comparison to the target reduction. Therefore, we proceeded to investigate methods which are effective at the expense of losing some information.

3.2.2. Nonspecific Within-Frame Coding. The codes considered in this section take advantage of the fact that not all the information in the original images has to be transmitted. There are two separable--but not independent--coding problems: the *image* code and the *transmission* code. We are concerned primarily with the image code, but we need to also specify a transmission code in order to evaluate the transmission requirement of the image code.

3.2.2.1. *Low-pass filtering (postage stamp experiment).* Among the reasons that allow us to transmit less information are the limitations of the human visual system and the fact that not all details are required for an observer to be able to recognize the message in the image. For example, in the "postage stamp" experiment described above, information is lost because of subsampling and low-pass spatial frequency filtering (imperfect resolution by the television display). A digital approximation to the analog "postage stamp" stimuli (with all stimuli scaled to the same size) is shown in Figs. 7a through 7d.

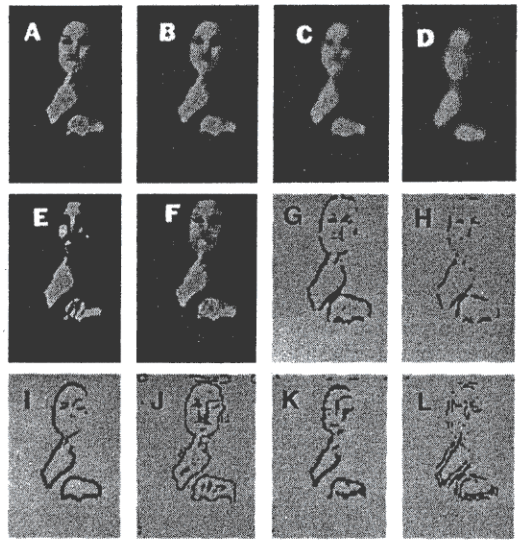

Fig. 7. Frame 20 of the "tomato" sequence, 64x96 pixels, subjected to 11 image transformations. The image coding scheme, and the average number of bits (1000 bits = kb) per frame in "tomato" sequence are indicated for each panel. The transmission code used to compute the bit rate is DPCM for frames (A-D) and hierarchical coding (Cohen [8]) for (E-L). (A) Original (18.4 kb). (B, C, D) Low-pass spatial filtering (8.7, 3.5, 1.3 kb). (E) Binary intensity, 6% black (0.9 kb). (F) Block truncation (3.7 kb). (G) Positive Laplacian, 6% black (1.3 kb). (H) Positive Laplacian, 3% black (1.0 kb). (I) Edge detection mask (1.3 kb). (J) Zero crossings 6% black (2.0 kb). (K) Negative Laplacian, 6% black (1.3 kb). (L) Union of positive and negative 3% black Laplacians (1.5 kb)

3.2.2.2. *Block truncation codes.* There are many other quite powerful methods for data compression at the cost of eliminating some (hopefully unimportant) information. Block truncation coding [29] is among the most efficient image codes that utilizes the full range of grey scale values (see Fig. 7f). In this coding scheme, each frame is divided into blocks that are transmitted separately. The code for a block depends upon the variance of the pixels inside the block. When applied to the ASL images, the code required approximately 0.6 bits per pixel for intelligibility, representing a net reduction of the amount of information by a factor of 20 relative to the original image. We plan to study an elaboration of the code in which blocks that change between frames are allotted more bits of information than unchanging blocks.

Even though the block truncation method is a quite elaborate code for full grey scale, the effective transmission rate for ASL is still 55,000 bits per second (96x64 pixels x 0.6 bits/pixel x 15 fps). This rate is many times higher, we suspect, than that essential for sign language communication. More complex methods are required, methods that are adapted, at least to some degree, to the idiosyncratic features of ASL communication.

3.2.2.3. *Binary intensity code.* The conversion of analog images into digital form without compression usually results in at least a five-fold increase in the required transmission bandwidth. However, the quantizing process also offers a means of compression by a reduction of the number of quantizing levels. The ultimate minimum number of intensity values in an informative image is two: black and white. Figure 7e shows a binary image. ASL sequences of these images are surprisingly comprehensible. In part, this is because the signer is wearing a black sweater and is viewed against a black background, so that the hands and face are the only nonblack image components. These conditions, obviously, are ideal for enhancing intelligibility of ASL, and particularly, of binary images of ASL.

Binary images lend themselves to run length encoding (for transmission) and to various hierarchical codes. The average number of bits per picture needed to encode these images is about 900. This is a coding efficiency of 0.15 bits per pixel, a 50 fold saving over the original's 7 bits per pixel. At 15 frames per second (approximately the minimum acceptable number for ASL), the required channel capacity is 13,500 bits per second (13.5 kbaud).

3.2.3. **ASL-Specific Within-Frame Coding - Cartoon Codes.** The important information in ASL is conveyed by the position of signer's arms, hands and fingers relative to the face and body. This information can be communicated by outlining the boundaries of the extremities. The resulting image is basically a cartoon-like representation of the signer.

3.2.3.1. *Edge detection by masks.* In order to determine the boundaries of objects, it is convenient to find most of the *edges* in an image. These are then connected to form closed curves. Fortunately, there has been considerable theoretical and empirical work done on the problem of edge extraction. By and large, the techniques applied involved various modification of gradient operators and masks. One method [35] selects the maximum of four different 3 by 3 masks applied at each point and retains the points with the largest resulting values. In the example of Fig. 7i only the largest 6% of the values were retained.

3.2.3.2. *Edge detection by Laplacian operators.* MARR and HILDRETH [27] proposed this method, but we are using several simpler variations on their algorithms that we have developed. Our preliminary observations indicated that a 5x5 linear operator applied to 64x96 pixel frames, followed by a thresholding to preserve only the strongest edges, was quite effective. Setting the threshold such that only 6% of the most prominent edge pixels were retained (Fig. 7g) results in a net reduc-

tion of the amount of information to 0.16 bits per pixel using a hierarchical coding scheme [8]. The case when only 3% of the edges were kept is illustrated in Fig. 7h. Detection of the positive side of edges is illustrated in Figs. 7g, h, the negative side of edges in Fig 7k, and the union of negative and positive edges in Fig. 7l. Positive and negative Laplacians reveal somewhat different aspects of images; both are about equally effective for this subject matter. The union is included to illustrate this.

3.2.3.3. *Zero crossings.* In addition to detecting the inside and outside of edges by the extreme values of the Laplacian operators, Marr and Hildreth proposed finding the midpoint of edges by locating the Laplacian zero crossings. We use a different algorithm to locate zero crossings; Figure 7j illustrates zero crossings based on the same Laplacian operator described above.

Most of the segmentation schemes based on edge detection require a method for combining the detected edges into a continuous boundaries. However, in the case of the sign language images, we found an interesting effect: Since the ultimate receiver is a human observer and the individual frames are presented in a rapid succession, the combination process occurs in the visual system of the observers. Even when each frame contains only disconnected edges, a sequence of frames appears to represent the boundaries quite effectively.[5]

The application of cartoon schemes followed by hierarchical coding enabled an information reduction from the original image by a factor of about 30. However, the resulting image still requires about 1300 bits/frame [8]. Intelligibility is maintained at frame rates as low as 15 frames per second, resulting in a transmission rate of about 20 kilobaud for 64x96 cartoon-coded ASL. The 20 kbaud rate for cartoons and the 13.5 kbaud rate for 1-bit intensity quantized images could be reduced somewhat by using a coarser pixel grid (fewer pixels) and by refinements in coding procedures to bring it below 10 kbaud, which is available on ordinary switched telephone networks. Substantial further reduction in transmission rate of binary images--or the transmission of grey scale pictures at this rate--may well require incorporation into the code of the physical constraints imposed by the construction of a human body, and the knowledge of ASL (signs, syntax and semantics).[6]

3.3. Information processing strategies

How is it that signers can do so well with such reduced visual information? In fact, their performance is not so different from other skilled human performance: listeners can understand speech in incredible amounts of noise; experienced military pilots have landed airplanes in virtually zero visibility; and chess masters can accurately reconstruct complicated chess positions from memory when their viewing time of the chess position is restricted to a few seconds.

The memory performance of chess masters provides perhaps the purest prototype for all these skilled performances. There are two studies of the memory of of chess masters for chess positions. DE GROOT [10] found that after a five second look at a chess board, his master could reconstruct it, correctly placing 24 or 25 of 25 pieces on an empty chessboard. CHASE and SIMON [5] found that their master could correctly place about 20 of 25 pieces from comparable positions. Their class A chess players could correctly place only half as many pieces as the master, and a novice could place only half as many again. The extraordinary performance of chess masters obtained only when the pieces were in the kinds of positions reached by skilled players in actual games. When chess positions were constructed by placing chess pieces randomly on the board, both the novice and class A player actually recalled the positions better than the master!

3.3.1. **Templates. Intentions.** The extraordinary memory of chess masters for chess positions (together with other aspects of their performance) has been interpreted in terms of what we here call memory *templates.* These presumed templates are memories for configurations of chess pieces that have been learned--stored in long term memory--during many thousands of hours of viewing chess positions. Obviously, templates are not literal representations of the chess pieces themselves, they are higher-order representations of chess configurations. In other contexts, it might be appropriate to call them *propositions.* It is estimated--but not known for certain--that templates typically deal with half a dozen pieces and that a chess master has acquired on the order of tens of thousands of templates. A master analyzes a novel position into its component templates, the deviations of particular pieces from these, and the connections between templates. It is reasonable to suppose, but not known, that templates themselves are organized into hierarchies of templates of templates.

We suppose that skilled ASL communicators have built up vast reservoirs of templates for hand and body configurations, and that speakers of a language have acquired sound templates of words and of typical sound sequences. Again, we suppose that such templates are represented in a more abstract feature space, where variations in the angle of view or individual differences between signers do not affect the representation.

3.3.1.1. *A gedanken experiment.* There is an additional difficulty in the study of language, relative to chess. Actual ASL signing and actual speech proceed so rapidly that the target positions of articulators are never quite reached. As a gedanken experiment, imagine a chess game being played so fast that the pieces never quite settle down. Before a piece being played by one player has actually touched the board, the other player has, perhaps, already begun to remove it and simultaneously has initiated the movement of another piece. We could simulate such a chess game by representing the chess board on a computer display. Players enter moves by keying a code at a terminal. This causes the displayed representation of the piece to move as though it were a massive body moving through an extremely viscous medium, taking many seconds to reach its destination. After a period of familiarization with the dynamics of the system, players could initiate new moves long before old ones had settled. In a rapid game between experienced players, photographs of the display board would never show all the pieces in resting positions and would be quite difficult to interpret. The gedanken chess board, we claim, represents the first-order situation in language communication, in which the target positions of the articulators--vocal tract or hands--are virtually never reached in actual discourse. Speakers and signers, like experienced chess players in the simulator, recognize the *intended* positions and reply in kind. In fact, in rapid discourse, speakers do not simply speed up--making the same articulatory gestures as in slow discourse at a faster rate; there are complex changes, including modifying their articulatory gestures so as to make their vocal intentions more discernible to the listener. The analogous modification of rapid movements probably occurs in ASL but it has not been reported.

3.3.1.2. *Intentions.* By *intention,* we mean an inference of the intended target position of a chess piece in the simulated game or of an articulator in spoken or signed language communication. To parse spoken or signed discourse we have to discover these intentions. According to this view, continuous discourse can be parsed into a sequence of target positions. Sequences of these target positions are represented as templates in the listener's memory, representing words or syllables. For ASL, the number of templates is presumed to be on the order of thousands, or at most, a few ten thousands. The reason a skilled signer can communicate with reduced visual input is that he does not attempt to parse the

input into an arbitrary sequence of movements, but uses it only as evidence for choosing among templates. The number of eligible templates is further constrained by context and pragmatics as has been pointed out above.

3.3.2. Analysis by synthesis, schemas. It is not necessary to be able to make signs in order to be able to understand ASL, nor is it necessary to be able to speak in order to understand speech. What is necessary is an understanding of the dynamics--the physical and biological constraints--of the articulators in order to derive the intention of a movement trajectory. A dog knows where to look for a stick when his master throws it or pretends to throw it. The dog can derive the intended trajectory from his master's arm and body movement even though a dog can not throw sticks. On the other hand, to utilize the constraints in a signed or spoken message requires a very sophisticated understanding of language, and of meaning. One proposed way of utilizing these constraints is to construct the alternative messages and to evaluate the incoming signals in terms of its evidential weight for these alternatives. The difficulties inherent in such procedures will be discussed in other papers here.

To represent meaning in a message SCHANK and ABELSON [36] and many others have proposed *schemas* as an appropriate means of representing meaning. A schema for representing a kitchen scene, for example, would include a floor, walls, cabinets, stove, refrigerator, utensils, a person, etc. A picture of a particular kitchen scene would be represented by filling details for the variables listed above, with default values representing the most typical scene. Evidence that people remember meaningful scenes in this way comes from their good recall of unusual details, and the nature of the errors they make, the errors being presumed to reflect default values. A schema can be viewed as a higher-order template whose components are variables, with values to be derived from the input.

To return to ASL, it is an article of faith that the representation of meaning of ASL sequences will not differ from the representation of meaning of spoken sentences. On the other hand, the surface form of ASL is quite different from spoken language; the grammar and semantics of ASL are quite different from spoken languages. Certainly, despite its name, American Sign Language bears little relation to English. The major advantages of studying a visible language like ASL is that the articulations--the hands and body--are continuously visible, and their physical properties can be readily measured. The inaccessibility of the articulators for speech, and hence the extraordinary difficulty of obtaining a dynamic description of their movements in speech [14] has been one of the stumbling blocks to an understanding of the speech process. ASL does not have this problem and seems ideally suited to the study of language processing at all levels by monkeys [44], or humans, or machines.

4. Acknowledgements

The preparation of this article and the work on motion perception was supported by U. S. Air Force, Life Sciences Directorate, Grant No. AFOSR-80-0279; the work on image processing of American Sign Language was supported by National Science Foundation, Science and Technology to Aid the Handicapped, Grant No. PFR-80171189.

The authors wish to acknowledge the technical, engineering assistance of Thomas Riedl and Robert Picardi, the advice of Dr. Nancy Frishberg, who along with Ellen Roth, also served as signer, and O. R. Mitchell, who made available his computer programs for block truncation codes.

5. Notes

1. When the two eyes are correctly pointed, we say they have achieved motor fusion. In this article, it will not be necessary to distinguish perceptual from motor fusion.

2. SPERLING [38] contains a fully detailed elaboration of the potential-theory model; see SPERLING [42] for a summary.

3. As defined, h represents a comparison in image space $I(x,y)$. As we shall observe later, comparisons (or template matches) are best made in feature space, after the most relevant, invariant stimulus properties have been abstracted. Thus h is better defined in terms of $T:I(x,y)$, where T represents the transformation to feature space ([38], p. 470; [26]), but T is an unnecessary complication at this juncture.

4. BRAUNSTEIN [2] proposes a similar theory (heuristics are used to disambiguate motion), and an informative review of motion cues.

5. The perceptual connection of dynamic segments into boundaries is analogous to JOHANSSON'S [20][21] classical demonstrations of the perceptual inference of moving objects (such as persons and bicycles) from several luminous points painted on their surfaces. POIZNER, BELLUGI and LUTES-DRISCOLL [33] tested the intelligibility of ASL communication when images were composed of lights placed on fingers, arms, and shoulders. Of the various arrangements tested, lights on the finger tips were most useful. TARTTER and KNOWLTON [43] demonstrated that ASL could be communicated by viewing 13 luminous points on each hand and one on the nose. In dynamic point-light displays, the reconstruction of hands occurs in the visual system of the viewer.

6. Since this article was prepared, the authors have become aware of two similar attempts to reduce the required transmission bandwidth of ASL by means of image encoding: (1) Pearson, D. E. & Six, H. Low data-rate moving-image transmission for deaf communication. *International Conference on Electronic Image Processing, 1982*, 204-208. (2) Abramatic, J. F., Leteiller, P., & Nadler, M. A narrow-band video communication system for the transmission of sign language over ordinary telephone lines. *Conference Report*, August, 1982.

6. References

1. Bellugi, U., & Fischer, S. A comparison of sign language and spoken language. *Cognition*, 1972, **1**, 173-200.
2. Braunstein, M. L. *Depth Perception Through Motion.* New York: Academic Press, 1976.
3. Braunstein, M. L., Andersen, G. J., & Riefer, D. M. Distance perception by monocular observers: Conflicting dynamic cues. *Investigative Ophthalmology and Visual Science*, ARVO Supplement, 1982, **21**, 273.
4. Burt, P., & Sperling, G. Time, distance, and feature trade-offs in visual apparent motion. *Psychological Review*, 1981, **88**, 171-195.
5. Chase, W. G., & Simon, H. A. The mind's eye in chess. In W. G. Chase (Ed.), *Visual Information Processing.* New York: Academic Press, 1973, pp. 215-281.
6. Cohen, Y. Measurement of Video Noise. *Technical Report, Human Information Processing Laboratory, N.Y.U.* February, 1982a.
7. Cohen, Y. The HIPL Picture Processing Software. *Technical Report, Human Information Processing Laboratory, N.Y.U.* July, 1982b.
8. Cohen, Y. Hierarchical coding of binary images. *Technical Report, Human Information Processing Laboratory, N.Y.U.* August, 1982c.

9. Crater, T. V. The Picturephone system: Service standards. *Bell System Technical Journal*, 1971, **50**, 235-269.
10. de Groot, A. *Thought and Choice in Chess*. The Hague: Mouton, 1965.
11. Dev, P. Perception of depth surfaces in random-dot stereograms: a neural model. *International Journal of Man-Machine Studies*. 1975, **7**, 511-528.
12. Falmagne, J-C. Random Conjoint Measurement and Loudness Summation, *Psychological Review*, 1976, **83**, 65-79.
13. French, N. R., & Steinberg, J. C. Factors governing the intelligibility of speech sounds, *Journal of the Acoustical Society of America*, 1947, **19**, 90-119.
14. Fujimura, O. Modern methods of investigation in speech production. *Phonetica*, 1980, **37**, 38-54.
15. Grossberg, S. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 1973, *52*, 217-257.
16. Grossberg, S. Competition, decision, and consensus. *Journal of Mathematical Analysis and Applications*, 1978, **66**, 470-493.
17. Helmholtz, H. L. F., von. *Treatise on Physiological Optics*, (3rd ed.), J. P. C. Southall (Trans.). Rochester, N. Y.: Optical Society of America, 1924. (Reprinted. New York, N. Y.: Dover Publications, Inc., 1962.)
18. Hillen, C. F. J. The face to face telephone. *Post Office Telecommun. Journal*, 1972, **24**, 4-7.
19. Johansson, G. *Configurations in Event Perception*. Stockholm, Sweden: Almqvist & Wiksell, 1950.
20. Johansson, G. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 1973, **14**, 201-211.
21. Johansson, G. Visual motion perception. *Scientific American*, 1975, **232**, 76-88.
22. Julesz, B. *Foundations of Cyclopean Perception*. Chicago: University of Chicago Press, 1971.
23. Klinger A., & Dyer C. R. Experiments on picture representation using regular decomposition. *Computer graphics and image processing* 1976, **5**, 68-105.
24. Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. *Foundations of Measurement*. New York: Academic Press, 1971.
25. Landy, M. The HIPL Picture/Header Format Standard. *Technical Report, Human Information Processing Laboratory, N.Y.U.* March, 1982.
26. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman and Company, 1982.
27. Marr, D., & Hildreth, E. Theory of edge detection. *Proceedings of the Royal Society of London*, **207**, 187-217.
28. Marr, D., & Poggio, T. Cooperative computation of stereo disparity. *Science*, 1976, **194**, 283-287.
29. Mitchell, O. R., & Delp, E. J. Multilevel graphics representation using block truncation coding. *Proceedings of the IEEE*, 1980, **68**, 868-873.
30. Nelson, J. I. Globality and stereoscopic fusion in binocular vision. *Journal of Theoretical Biology*, **49**, 1-88.
31. Pearson, D. E. *Transmission and Display of Pictorial Information*. New York: Wiley, 1975.
32. Pearson, D. E. Visual communication systems for the deaf. *IEEE Trans. on Communications*, 1981, **29**, 1986-1992.

33. Poizner, H., Bellugi, U., and Lutes-Driscoll, V. Perception of American Sign Language in Dynamic Point-Light Displays. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, **7**, 430-440.

34. Restle, F. Coding theory of the perception of motion configurations. *Psychological Review*, 1979, **86**, 1-24.

35. Robinson, G. S. Edge detection by compass gradient masks. *Computer Graphics and Image Processing*, 1977, **6**, 492-501.

36. Schank, R. C., & Abelson, R. P. *Scripts, Plans, Goals and Understanding*. Hillsdale, N. J.: Lawrence Erlbaum Assoc., 1977.

37. Schwartz, B. J., & Sperling, G. Nonrigid perceptions consistently elicited by rigid object stimuli. Manuscript submitted for publication, 1982.

38. Sperling, G. Binocular vision: a physical and a neural theory. *American Journal of Psychology*, 1970, **83**, 461-534.

39. Sperling, G. Future prospects in language and communication for the congenitally deaf. In L. Liben (Ed.), *Deaf children: Developmental perspectives*. New York, N. Y.: Academic Press, 1978. Pp. 103-114.

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

41. Sperling, G. Video transmission of American Sign Language and finger spelling: present and projected bandwidth requirements. In A. Habibi and A. N. Netravali (Eds.), *IEEE Transactions on Communication* [Special Issue on Picture Communication Systems]. New York: IEEE Communications Society, 1981a, 1993-2002.

42. Sperling, G. Mathematical models of binocular vision. In S. Grossberg (Ed.), *Mathematical Psychology and Psychophysiology*. Providence, Rhode Island: Society of Industrial and Applied Mathematics-American Mathematical Association (SIAM-AMS) Proceedings, 1981b, **13**, 281-300.

43. Tartter, V. C., & Knowlton, K. C. Perception of sign language from an array of 27 moving spots. *Nature*, 1981, **289**, 676-678.

44. Terrace, H. S. *Nim*. New York, N. Y.: A. Knopf, 1979.

45. Todd, R. Visual information about rigid and nonrigid motion: A geometric analysis. *Journal of Experimental Psychology, Human Perception and Performance*, 1982, **8**, 238-252.

46. Ullman, S. *The Interpretation of Visual Motion*. Cambridge, Massachusetts: The MIT Press, 1979.

47. Technical Staff, Bell Labs. *Transmission Systems for Communications* (Rev. ed. 4). Western Electric Co., Winston-Salem, N. C. Tech. Pub., 1971.