Perception & Psychophysics
1973, Vol. 14, No. 1, 90-94

Scanning visual images: Some structural implications

STEPHEN MICHAEL KOSSLYN*
Stanford University, Stanford, California 94305

This experiment was designed to explore the spatial and structural properties of visual imagery. Forty Ss were shown drawings and later asked to verify pictorial features of the drawings from memory. One group of Ss was instructed to be able to recall an image of each drawing and to focus initially on one specified end of their images during the subsequent verification task. Another group of Ss were asked to recall a verbal description of each drawing and initially to describe one specified end of a drawing during the verification task. Time to verify pictorial properties was a function of the spatial distance of a property from an initial focus point for both groups, but Ss in the verbal description group experienced much greater difficulty in performing the task. Comparison of these times with those from additional imagery encoding and verbal encoding groups given no focusing instructions indicated that focusing instructions effectively directed scanning strategies.

The study of mental imagery has taken two basic forms in recent years. Most research has been *functionalist* in orientation; that is, interest has centered primarily on the role or function of imagery as a mnemonic (e.g., Bower, 1972; Paivio, 1972), cognitive strategy (e.g., Huttenlocher & Higgins, 1972), or analogue to perception (e.g., Segal & Fussella, 1970, 1971). Studies investigating structural properties of the image itself, on the other hand, have been infrequent. Shepard and Feng (1972) performed one of the few studies that have been *structuralist* in orientation. They presented Ss with pictures of unfolded cubes that had an arrow drawn on two of the squares (unfolded sides). The task was to "mentally fold" the cubes and to indicate as quickly as possible whether or not the sides the arrows pointed to would be adjacent (would meet) in the folded cube. Verification times were found to be linearly related to the number of folds necessary to join the arrows. Similarly, Cooper and Shepard (1972) asked Ss to determine whether alphanumeric characters, which were presented at one of several orientations about the circle, were normal or mirror-image versions. Verification times were a function of the angular departure of the test character from the standard upright position. Presumably, these times reflect the amount of time to "mentally rotate" an image of the test character to the upright position. These and other studies (cf. Cooper & Shepard, 1972) demonstrate that some types of visual images not only maintain internal structures analagous to those of the referents, but also exhibit some of the same spatial properties as do the original percepts.

The notion that the internal structure of an image parallels the spatial structure of its referent might generally characterize long-term memory, or "generative," images. This is the kind of imagery Galton studied and poets hope to evoke; it is also the imagery most often studied by psychologists. If generative images reflect the spatial organization of the items imaged, as Neisser (1970) suggests, we might expect several things. First, people ought to be able to focus selectively on a part or feature of an image, just as they can attend to selected portions of pictures (even when no eye movements are possible). Second, we might also expect that time would be required to shift attention from one part of an image to another. Finally, if images are indeed spatial, then the farther apart one property is from another, the longer it should take to shift attention from that property to the other. The present study tests these hypotheses.

An investigation of imaginal phenomena must include controls for possible verbal-linguistic confoundings and artifacts. Many "imaginal" processes can, in fact, be adequately accounted for in terms of propositional encodings (e.g., Clark, 1972). The present study attempted to anticipate what nonimaginal verbal strategies Ss could possibly use and then to create special control groups explicitly instructed to use that strategy. A verbal-encoding interpretation of the finding that time to verify "properties of an image" increases as spatial distance from an initial focus point to a feature increases might go something like this: When S originally encounters a picture (for example) he describes it to himself. Propositions formed contiguously as S is looking at the picture form adjacent items on a list. Increasing time to verify features requiring "scanning larger distances on an image" would be accounted for by positing more intervening propositions to be searched through on a list.

It is, however, possible to distinguish this kind of propositional representation from an imaginal organization. There is no a priori reason to expect that it would be any more difficult to scan an image in any one given direction than in any other direction. A list of propositions, on the other hand, might be characterized

*The author conducted this research while receiving support from an NSF graduate fellowship (1970-72). The author wishes to thank Eleanor Maccoby for use of her tape recorder, purchased under NIH HD 00125, Peter Lucy, Tom Schumacher, Tom Roberts, and Joyce Lockwood, for their technical assistance, and Herbert Clark, Edward Smith, John Anderson, Bill Banks, Gordon Bower, Lynn Cooper, and Caroline Bowker, for their invaluable suggestions and advice.

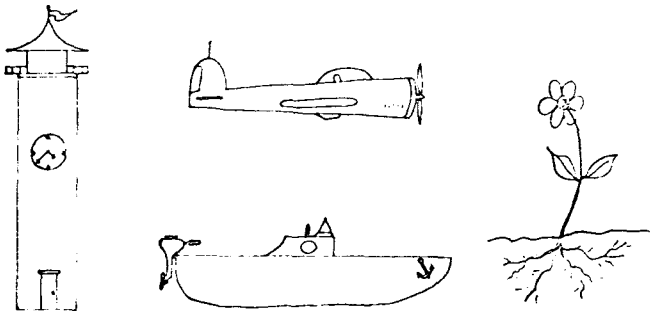

Fig. 1. Examples of line drawings used.

by asymmetrical associations; going "down" the list may be easier than going "up" it, for example. Such asymmetrical associations between list items seems especially likely for representations of horizontal objects where strong left-to-right encoding biases (due to reading habits) probably exist. Furthermore, the order in which propositions about an object were formed need not exactly mirror spatial properties of the object. Thus, if Ss are instructed to encode visual stimuli propositionally, one might expect greater difficulty in retrieving various spatial features from preset spatial locations (requiring a scanning search) than if imagery is used.

The present experiment, then, is concerned with: (1) whether or not Ss can selectively attend to portions of an image, (2) whether or not time to retrieve features of an image is a function of spatial distance scanned during search, and (3) if imaginal representations of visual stimuli can be distinguished from verbal representations.

METHOD

Ss were shown pictures and asked to be able to recall an image or description of each one. Ss then participated in a reaction-time (RT) task requiring verification of properties of the pictures from memory. Half of the Ss in each encoding mode were instructed to "focus" attention on a specified end of their picture-representation before probe properties were presented and half of the Ss were instructed to keep the whole picture in mind.

Materials

Ten line drawings of common objects were prepared (see Fig. 1 for examples). All objects were oblong, half being vertical and half being horizontal. The vertical objects pictured were: a tower, a plant, a man, a rocket, and a lamp; the horizontal objects were: a car, a speedboat, a rattlesnake, a plane, and a locomotive. All drawings had some easily labelable features at both ends and midway between the ends.

A tape recording was prepared such that the name of a drawing (e.g., "car") was followed 5 sec later by a possible property (e.g., "headlight"). Between each property and the following name was 10 sec of silence. Each of the 10 names of drawings was paired with six property words, three of which described actual features of the drawing ("true" properties) and three of which did not ("false" properties). Of the true properties, one referred to a feature at the left or bottom of the drawing (depending on whether it was a horizontal or a vertical drawing), one referred to a centrally located property, and one was a rightmost or top-located property. The three true property names for each object were all either one or two syllables long. For example, for "speedboat," the "true" properties were "motor," "porthole," and "anchor." The names of each drawing were randomized and then recorded in the same sequence six times in succession to minimize any possible effects of different amounts of interpolated material on image production. The six selected property words were paired randomly with the appropriate name at any given iteration (i.e., the first through the sixth time the name was recorded).

Procedure

Instructions were prerecorded; which recording a particular S heard was a function of which of the four experimental groups he was randomly assigned to. Ss in Group 1, the "whole imagers," were shown the drawings and told that they should attempt to remember the name of each drawing (provided during presentation) and what each drawing looked like well enough to generate an accurate visual image of it. Ss in this group, as well as those in Group 2, first saw each drawing for 10 sec. Following this, E went through the drawings again, but this time S heard the name 5 sec before actually seeing the drawings. During this initial 5 sec, S was to make an image of the named drawing and was then to use the second presentation to correct and improve his memory. Following this, S participated in the RT task. Upon hearing the first word of a pair (a name of a drawing), S was to make an image of the *whole* drawing. S was instructed to "look on" his image when he heard the second word of a pair (a possible property) and to indicate, by depressing one of two buttons as quickly as possible, whether he "saw" the property on his image.

Ss in Group 2, the "focus imagers," received imagery encoding instructions and procedure identical to those used for Group 1. The only difference between Groups 1 and 2 was in the instructions for the RT task. Ss in Group 2, in contrast to those in Group 1, were told that upon hearing a drawing's name they should make a visual image of the appropriate picture and *focus* their attention on one prescribed end as if they were "staring at that place on the picture." When the property word was presented, those Ss were to cease focusing on the specified end and to "look for the property" in the same manner as were Ss in Group 1. It was emphasized that S should wait for the property word to be presented before ceasing to fixate mentally on the preset focus point. Half of the Ss in this group were instructed to focus on the extreme left if the drawing was horizontal or on the bottom if it was vertical; the remaining half focused on the right or top ends of their images.

The third group, the "whole verbalizers," did not receive imagery-encoding instructions. This group was told that they would be shown pictures and should describe the pictures silently to themselves and later be able to recall what description went with each drawing's name (given by E during presentation of the pictures). Following these instructions, S was shown the drawings one at a time and given 10 sec with each of them. After this, E spoke each drawing's name 5 sec before showing S the picture again. In the 5 intervening seconds, S was to describe the drawing to himself, and then to compare his description with the actual picture, making necessary corrections and improvements. Ss in this group finally were instructed to begin describing the *whole* drawing to themselves as soon as a name was presented during the RT task. Upon presentation of the property word, S was to check through his description for that property and to respond by depressing the appropriate button.

The fourth and final group, the "focus verbalizers," had encoding instructions identical to those of Group 3. The only difference between this group and Group 3 was in the instructions for the RT task. These Ss were told that upon hearing the drawing's name they should begin describing a

specified end (half, the left or bottom, the other half, the right or top) of the picture until the property word was presented; only after the property word came on should S think of anything other than the description of the prescribed end. As in Group 3, Ss were to check their descriptions when verifying properties. Postexperimental debriefing indicated that the Ss in Groups 3 and 4 usually encoded the target properties. Description lists often appeared hierarchically organized; that is, for "speedboat," for example, "motor" might have been recalled as well as "handle," "propellor," "pulley on top of motor," etc.

Half the Ss made "true" responses with their right hands, half with their left. The Ss were instructed to respond as quickly as possible, keeping errors to a minimum. Before hearing the tape, S was asked to reiterate the instructions. Any misconceptions were corrected and emphasis was put on following the instructions exactly; focus groups were reminded again to focus on the directed part of a drawing in their description or image. The S was then told that before the actual test items came on there would be a few practice items to make sure he understood the task. The item names in the practice trials were of common objects and S was to image or describe, as the case might be, any particular instance he chose of the object. After the eigh practice name-property pairs, S was asked what he did upon hearing the name of each object. Once again, the directions for the S were reiterated and emphasis was placed on following instructions as well as possible; pilot work had indicated that considerable instructional overkill was necessary to insure Ss' compliance. The S then heard the actual test items. Upon presentation of each property word, an electronic clock was initiated. E, his face concealed from S, recorded RTs and truth judgments. After the last item, E inquired as to "what went on in your mind when you heard the last object named?" The S often did not realize that he was being indirectly asked if he had followed instructions, and a number of Ss gave replies widely differing from the original instructions. In such cases, S was asked how much of the time he used the alternate strategy (almost always whole imagery); if it was greater than about one-third of the time, S's data was discarded.

Subjects

Of the 49 Ss tested, from an introductory psychology course population, only 40 were actually included in the analysis (10 Ss in each group). Four Ss in the focus-imagery, 3 in the focus-verbal, and 2 in the whole-verbal groups indicated that they did not follow instructions at least one-third of the time when queried immediately after the session.

RESULTS

Three mean RTs to "true" properties were obtained for each S. For "focus" groups, RTs were analyzed in terms of distance of verified property from the focal point; a mean was calculated for items located at the point of focus, located in the middle, and located at the opposite end of the drawing. For "whole" groups, RTs to the leftmost items on horizontal pictures were grouped with RTs to items on the bottom of vertical drawings; all middle items were grouped together, as were right and topmost pictorial properties. This was the same combination that was used in the focus groups, providing a control where the effects of instructions on the same items could be compared. The "whole" data from horizontal and vertical images and descriptions could have been grouped in the reverse manner (left-top,

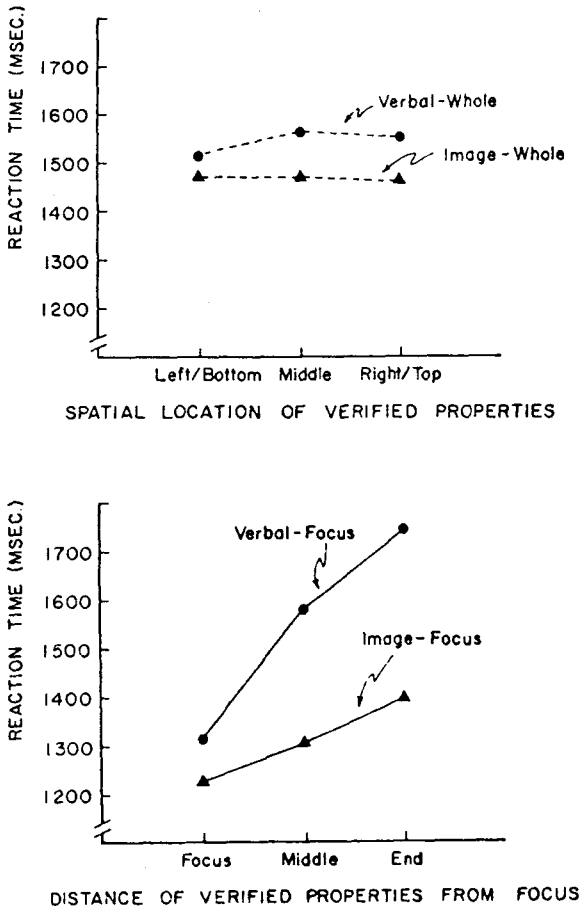

Fig. 2. Speed of verification of picture properties as a function of spatial location and distance from the point of focus.

right-bottom) without affecting the results within these groups, as there was no significant difference in RTs to properties from different spatial locations on horizontal [$F(2,36) < 1$] or on vertical pictures [$F(2,36) < 1$].

The basic predictions of the study were confirmed (see Fig. 2). The fact that people can selectively attend to portions of an image was revealed by the significant advantage of focusing at the general spatial location of a queried property. This advantage was reflected in the significant interaction between focusing or whole instructions and RTs to properties from various spatial locations [$F(2,36) = 6.39$, $p < .01$]. Not only were properties at the point of focus fastest for the focus-imagery group, but time to verify properties at other locations was a function of how far they were from the point of focus [increase in RT at opposite end, 168 msec; linear trend, $F(1,36) = 4.75$, $p < .05$]. Also as predicted, the effects of focusing were much more pronounced for the verbal encoders [399-msec increase; linear trend significant at $F(1,36) = 26.78$, $p < .001$]. The fact that time to retrieve information from spatial locations away from the point of focus increased more sharply for verbal encoders than for imagery encoders

was further attested to by a significant interaction of the linear components of slopes from focus-verbalizers and focus-imagers [$F(1,36) = 4.49$, $p < .05$].

The possibility of assymetrical associations in a list was cited as one reason the focus-verbalizers might experience more difficulty in the task than the focus-imagers. This notion seems especially applicable to encodings of horizontal items where highly overlearned left-right encoding biases might be expected. We hypothesized, then, that data from horizontal drawings might further reflect differences between strategies used by the focus-imagers and the focus-verbalizers. This expectation was confirmed; not only did horizontal items take longer to respond to in general [$F(1,36) = 4.24$, $p < .05$], but most of this effect was due to the verbalizers [as indicated by a significant interaction between strategy type and drawing orientation, $F(1,36) = 7.99$, $p < .01$]. Moreover, this result was not simply a consequence of verbal encodings somehow being more difficult in general. Rather, it was an outcome of greater difficulty in retrieving information from selected spatial locations. The increase in RTs to properties at the opposite end of a drawing from those to properties at the focus point was greater for verbalizer horizontals (493-msec increase) than for imagery horizontals (184-msec increase); the linear components of these slopes did interact significantly [$F(1,36) = 5.24$, $p < .05$], although none of the other possible comparisons of trends (e.g., imagery verticals vs verbal verticals, etc.) did so. The hypothesis that scanning "up" a list would be more difficult than scanning "down" a list seemed to be supported by the finding that scanning for features located to the right of the focus point in general was faster (139-msec increase) than it was in the reverse direction [539-msec slope; interaction of linear components, $F(1,32) = 18.21$, $p < .001$]. In order to assess how much of this effect was due to verbalizers, it was necessary to analyze ease of scanning in both directions on horizontals separately for imagers and verbalizers (a slightly questionable procedure, as each data point has a maximum of 25 observations from only five Ss). The outcome of this analysis is clear-cut. The focus-verbalizer group scanning right to left exhibited a very large (799-msec) increase, while the imagery group scanning in the same direction was markedly quicker (290-msec difference). The interaction of the linear components of these slopes was highly significant [$F(1,32) = 14.11$, $p < .001$]. In contrast, the interaction of the left-to-right scanning slopes from the two representational modes (imagery = 79 msec, verbal = 199 msec) was not significant [$F(1,32) < 1$]. Furthermore, the prediction of unequal ease of accessing from different ends of a list was supported by the significant left vs right focus-point interaction [$F(1,32) = 9.87$, $p < .005$] in the focus-verbalizer data. The focus-image data, on the other hand, did not exhibit a significant interaction as a function of accessing from left vs right [$F(1,32) = 1.27$, $p > .10$].

Although scanning top to bottom for properties of vertical pictures was faster (161-msec increase) than in the opposite direction [379-msec increase; interaction of linear components significant at $F(1,32) = 5.16$, $p < .05$], none of the effects of representational mode noted above obtained. This may have been because Ss in the focus-verbalizer group did not uniformly encode in a top-to-bottom manner, but also encoded in the opposite direction as well. This would result in the associations between propositions being more symmetrical than those formed when horizontal pictures were encoded.

In closing, it should be noted that none of the between-S main effects were significant [image vs verbal, $F(1,16) = 1.01$; focus vs whole, $F(1,16) = 2.90$, $p < .10$]. Similarly, the difference in RTs to items located in the middle of a picture for the focus-imagers and the whole-imagers (see Fig. 2) was not significant [$F(1,36) = 2.41$, $p > .10$]. These RTs should be about the same if Ss in the whole imagery group scanned, on the average, about half of an image per trial, as would be expected if the distance between spontaneous focus points and probed features was random.

The overall error rate was 9%. Errors and wild scores were not included when calculating the means. "Wild" scores were defined as those that exceeded twice the mean of the remaining scores in a cell; only one score per cell could be so eliminated.

DISCUSSION

The results clearly indicated that people can selectively retrieve information from preset spatial locations on a generative image. Furthermore, retrieval of items from an image is a function of actual physical distance from the point of initial focus. One way to look at this result would be to think of an internal representation analogous to the actual picture where S merely fixates on one part, and then scans over the representation if the queried feature is not at the point of focus. Another way of interpreting this finding, which is consistent with some S's introspections, is to think of S having an image of only the part he is immediately "looking" at. When the query comes on, he then retrieves the remainder of the image from memory. In the first case, the image is like a billboard that is all lit up at night and one merely stares at a selected portion of it. In the second case, the billboard is dark except for a portion that is under the immediate spotlight of attention. Both notions involve S's retrieving perceptual features from memory which are organized in terms of spatial relations. In the first case, images are conceived of as inherently integral—the parts are inextricably part of the whole. In the second case, images themselves may be retrieved piecemeal.

Given that an image might be a collection of features that may even be accessed serially, one might be tempted to argue that there is no real basis for distinguishing imaginal representations from purely

verbal modes. The data, however, do not support this contention. In the group given explicit instructions to use a verbal strategy, the slope was much steeper than the corresponding scan slope for the focus-imagery group. This additional steepness was largely a result of retrieval from descriptions of horizontal drawings, specifically in cases where right-to-left scanning was required. The underlying representations engendered in the focus-verbalizer group may not always have been spatially organized, which greatly hindered retrieval in the essentially spatial task. Furthermore, even when the organization of verbal material mirrors spatial organization, the associations between propositions need not be symmetrical, as is implicated in the strong interaction of left-right and right-left scan slopes in the focus-verbal group. Imagery representations, on the other hand, seemed to allow easier access from any point of entry.

An essential difference between imaginal and verbal modes of representation, then, would seem to lie in the intrinsically spatial nature of the image; it is as if all associations between properties are implicit within the context of the whole. This is not to claim that an image is entirely "integral" or "holistic"—errors are made, people do forget individual features of an image. The visual-spatial qualities of visual imagery imply that an imaginal "feature" ought to be distinguished from a verbal feature in that it is essentially perceptual, it is a "remembered appearance." A verbal encoding, on the other hand, is a step further removed—it is a proposition about an appearance (cf. Bower, 1972). This notion, that the features of an image are in some way isomorphic to "perceptual features" (cf. Gibson, 1967), might help to account for the hindering effects of imaging on like-modality perception (cf. Segal & Fussella, 1970). It is interesting, in this context, to note that the Ss in the imagery groups seemed to spontaneously shut their eyes during the RT task much more often than did the verbal encoders. Unfortunately, no accurate records were kept of this behavior during the task.

REFERENCES

Bower, G. H. Mental imagery and associative learning. In L. Gregg (Ed.), *Cognition in learning and memory.* New York: Wiley, 1972.

Clark, H. H. More about "adjectives, comparatives, and syllogisms": A reply to Huttenlocher and Higgins. Psychological Review, 1972, 78, 505-514.

Cooper, L. A., & Shepard, R. W. Chronometric studies of the rotation of mental images. Paper presented at the Symposium on Visual Information Processing, Carnegie-Mellon University, May 18-19, 1972.

Gibson, E. J. *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts, 1967.

Huttenlocher, J., & Higgins, E. T. Adjectives, comparatives and syllogisms. Psychological Review, 1971, 78, 487-504.

Neisser, V. Visual imagery as process and as experience. In J. S. Antrobus (Ed.), *Cognition and affect.* Boston: Little, Brown, 1970. Pp. 159-178.

Paivio, A. *Imagery and verbal processes.* New York: Holt, Rinehart & Winston, 1971.

Segal, S. J., & Fusella, V. Influence of imaged pictures and sounds on detection of visual and auditory signals. Journal of Experimental Psychology, 1970, 83, 458-464.

Segal, S. J., & Fusella, V. Effects of six sense modalities on detection of visual signal from noise. Psychonomic Science, 1971, 24, 55-56.

Shepard, R. W., & Feng, C. A chronometric study of mental paper folding. Cognitive Psychology, 1972, 3, 228-243.

(Received for publication November 22, 1972;
revision received March 13, 1973.)