ON RETRIEVING INFORMATION FROM VISUAL IMAGES*

Stephen Michael Kosslyn
The Johns Hopkins University
Baltimore MD

What shape are a German shepherd´s rs? By way of developing a set of common tuitions, it would be appreciated if the ader would answer this query. Most people port that in order to answer this question nswer: pointed), a visual image of the g´s head was generated and the ears spected. The present paper is concerned th this phenomenon, with how information represented in, and later extracted from, sual images. The present theoretical amework is in part presented in Kosslyn n press), which unfortunately is not ailable for public consumption at the time this writing. Thus, it seems valuable at esent to reiterate briefly these ideas in e course of further developing them.

Theoretical Framework

A computer graphics metaphor: A visual age is considered to bear the same basic lationship to its underlying structure as pictorial display on a cathode ray tube RT) does to the computer program that nerates it. The underlying "deep" ructure is abstract and not experienced rectly, whereas the image itself seems ctorial in nature. We are not claiming, wever, that the psychological analogue to e CRT displays *pictures* as such; rather, is structure is characterized as pporting internal representations hatever they may be like) similar to those at engender the experience of perceiving a cture when a person is viewing one. Thus, image of a triangle need not be reflected anything actually triangular occurring in e brain.

The mind´s eye: Once an image is nstructed, we claim it then may be terpreted in terms of various conceptual tegories. Following our metaphor, if we ink of an image as a design like that on a T, the "mind´s eye" may be regarded in rms of sets of processes that detect tterns (e.g., lines, angles, closed areas various sorts, etc.), and relationships tween patterns, in this display. This rt of classification may utilize a series "procedures" (in Winograd´s 1973, 1975, nse) that test for the criteria associated th membership in given conceptual tegories (like "pointedness," or "leg"). e present claim is that the same ocedures may be used appropriately for assifying both internal representations ising during perception which engender perience of a visual percept, and internal presentations experienced as a visual ntal image.

An image space: The psychological analogue to the CRT in our metaphor can be thought of as having spatial boundaries, although not necessarily ones as rigidly defined as are the edges of a CRT; mental images, many claim, can only be expanded in subjective size to a certain point before they seem to "overflow." We should note, however, that the display mechanism is inextricably linked to the processor, the "mind´s eye," that operates on its contents. It could be argued that spatial constraints imposed on subjective size of images are due to the "scope" of this processor. At the present juncture, the size of the display mechanism and the spatial extent available for further analysis are indistinguishable from each other; we may just as well consider material outside the processor´s scope as having overflowed an internal display screen.

Capacity limitations: Our metaphor suggests two loci where capacity limitations might occur: 1) Perhaps only a limited number of underlying units (which represent parts of an image), or limited proportion of the underlying structure, may be able to be activated simultaneously; 2) In addition, perhaps the analogue display itself is limited in terms of how much area or detail can be represented at any one time. Consider an analogy to a CRT display that requires a figure to be "refreshed" (re-painted) repeatedly by an electron gun in order to be maintained. That is, a stream of electrons is directed at a point behind the tube, and causes the phosphor coating on the screen at that point to glow. The stream is then moved elsewhere, or terminated, and the phosphor begins to fade; unless this point is again exposed to the electron beam, it soon will cease to glow altogether. If a displayed picture is composed of too many points, they will begin to fade noticeably ("flicker") between exposures to the electron beam.

In addition to being simpleminded, it undoubtedly is wrong to consider a human´s visual image in terms of a sequence of points, let alone as a sequence of points individually plotted. A better way to conceptualize imagery might be in terms of hierarchical representations. Subroutines for displaying lines, arcs, and a set of basic patterns might *serve* as primitives; these basic units could be built into the system or learned. All that need be stored in a particular perceptual memory is the highest levels of description, which can be used to direct the low-level subroutines to plot a display of the represented entity. Such notions might help us to understand why images seem composed of parts which never are only "half-present." A hierarchical conception also is more reasonable vis-a-vis storage requirements; having to store the locations of each and every segment of a figure would soon place excessive demands on memory as one accumulated image encodings. Thus, limits in the capacity of the analogue display mechanisms might best be thought about in terms of how many surface *parts* of an image can be displayed simultaneously

----

[wish to thank Steve Palmer and Jim merantz for helpful criticism and ggestions. Some of the work described rein was accomplished with the aid of omedical Sciences Grant 5 S05 RR07041-09 arded to the Johns Hopkins University by he Division of Research Resources, DHEW.

before imaging resembles a "mental juggling act," with pieces slipping in and out.

Miscellaneous implications: The basic computer graphics metaphor provides intuitively appealing (to many), ways of viewing many other aspects of imaging. The "vividness" of an image can be considered to depend on (1) how much detail is stored in the underlying structure and represented on the image, and (2) how often the parts of an image are refreshed from long-term memory. The experience of "focusing" on part of an image may reflect selective allocation of display capacity to a particular portion of an image. "Scanning" an image, then, would involve sequential focusing which shifts serially and continuously across an image. Many other imagery phenomena are easily thought about in this general framework; this exercise, however, soon loses its appeal as it becomes increasingly apparent that greater explicitness and specificity regarding the human system is required in order to conceptualize some of the more interesting imagery findings and introspections (e.g., mental rotation of images). Rather than fill out our model by making more-or-less arbitrary decisions, it seems most propitious to discover how useful the model is, as developed thus far. In the course of exploring empirical implications we should encounter data that will help us determine how best to make our conceptualizations more explicit and rigorous.

Some Experimental Findings

Basic methodology: Implications of the computer graphics metaphor have been explored primarily with the use of one sort of methodology. Subjects typically are asked to construct an image according to some specification (regarding size and/or context, depending on the experiment). Some seconds thereafter, a possible characteristic of the imaged animal, object and/or scene is presented. The subject is to evaluate, as quickly as possible, whether or not that characteristic is appropriate for the probed image. For example, if he is asked to image a lion and then hears "feet", he is to look on his image for feet. If the object imaged does in fact have the probed property, he is told, it ought to be locable on the image. Upon finding the feet, the subject responds "true" by pushing a button; if the property is not appropriate for the imaged entity, he depresses another button to respond "false." The measure of interest is the time required to evaluate his image. It is emphasized that this is an "internal detection task" where we are interested in how long it takes to use the image, to "see" parts, and we are not interested simply in how quickly a person can make the correct assessment by any available means.

The effects of size: As objects become smaller, their constituent parts become more difficult to discern perceptually. Many factors contribute to this effect. One factor may be that the procedures that categorize perceptual information need a minimal spatial extent upon which to operate efficiently. If so, our metaphor leads us to expect that properties of relatively small images also should be more difficult to identify. And in fact this is true. Kosslyn (in press) reports six experiments where parts required more time to identify when the object was imaged subjectively small compared to when the object was imaged subjectively larger. This result was obtained when people adjusted the subjective size of a single image (of an animal) directly, and when size was manipulated indirectly. In the latter cases, size of a probed image was manipulated by asking the subject to image the target object adjacent to a much larger or smaller object. For example, when subjects imaged an elephant and rabbit together -- correctly proportioned -- the rabbit was smaller than when it was instead imaged next to a fly. In this situation, more time was required to determine that a rabbit has ears, by referring to imaged rabbit, when it was adjacent to the elephant. In another experiment, this result was reversed simply by asking subjects to image target animals either next to huge flies or tiny elephants.

These results have been interpreted in terms of the parts of larger images themselves being larger. We would expect, by this logic, that larger parts -- in and of themselves -- ought to be detected more easily on a given image than smaller parts. Kosslyn (1974) reports an experiment where probed parts of animals were selected such that the smaller part (e.g., "claws" for cat) was more strongly associated with the animal in question than the larger property ("head," for cat). This situation allows one to distinguish between imagery use and accessing abstract representations of the sort described by Collins and Quillian (1969). If imagery is used, the larger part ought to be evaluated more quickly. If imagery is not used, not the smaller -- but more strongly associated -- part should be judged appropriate more quickly (c.f. Conrad, 1972; Smith, Shoben, & Rips, 1974). These results were obtained in all 3 conditions tested (which differed in terms of when a subject generated the inspected image relative to the time of probe).

The effects of load: Another phenomenon suggested by the computer graphics metaphor is that as more parts are added to an image, it ought to become more degraded due to capacity limitations. More time ought to be required to identify parts of more degraded images. And in fact this seems to be true: An animal property is evaluated more quickly when the animal is imaged next to a 4-cell matrix (that is, imaged as if it were painted on a wall to the animal´s left) than when the animal is imaged next to a 16-cell matrix. Similarly, properties of animals imaged next to images of two digits (painted on an imaginary wall) were evaluated faster than when four digits were in the image. In this situation, questions about the presence of particular digits in the image also were answered more quickly when two rather than our digits occurred. By varying the elative size of an animal and the number of ells in the matrix (or number of digits) maged next to it, it was demonstrated that ize and number of parts had independent ffects on time to verify properties of the mage (see Kosslyn, in press; 1974).

Time to generate images: The present etaphor suggests that images are onstructed, that the constructive processes onsume time, and that subjectively larger mages display more identifiable details han smaller images. These notions are upported by the results of a simple xperiment: Subjects were asked simply to ndicate when they had in consciousness a ental image of a given animal at one of our possible relative subjective sizes. As he subjective size increased, so did the ime necessary to construct the image. resumably, greater spatial extents take onger to fill in. Because these extents re greater, however, properties of ubjectively larger images are more readily ecognizable and detectable than properties f subjectively smaller images.

The maximal size of images: The present etaphor suggests that images cannot be xpanded in subjective size indefinitely efore seeming to "overflow". Kosslyn & alph (in preparation) tested one mplication of this idea. If one moves an bject in an image "closer," it will seem to xpand subjectively until it finally no onger is all in view at one time. If size s constrained, then images of smaller bjects ought to be able to be moved ubjectively closer prior to overflowing. hus, for example, a mouse ought to be isualized in its entirety at a much closer ange than an elephant; if the elephant were maged equivalently close one might "see" nly a patch of gray hide. This in fact was he result obtained in 5 experiments where eople estimated (in various ways) the istance of imaged objects at the point of verflow. The correlation between the ongest axis of the imaged object (which ccounted for the lion´s share of the ariance in a regression analysis) and stimated distance ranged from .86 to .98. e presently are using these data to measure he "internal visual angle" of the mind´s ye (if you will), hoping that it remains elatively constant when one images ifferent sized objects as large ubjectively as possible.

Erasing and replacing visual images: ur metaphor obviously does not commit us to ell-specified positions on all issues oncerning imagery. The phosphor which lows and allows display on a cathode ray ube may fade relatively quickly or slowly, epending on the variety. For human magery, we would expect the psychological nalogue to this display mechanism also to ave fade characteristics. Since images do equire effort to display internally, it ould seem economical to have a mechanism esigned to maintain them for a non-trivial amount of time thereafter. Thus, we might expect that images, once formed, take time to fade. During this "rebound period" new images should be difficult to construct. Smaller, or less area consuming, images should be easier to replace with a subsequent image because more "blank" space is available in which to begin immediate construction of the new image. With larger images, in contrast, the existent image must fade before a new one may begin to be constructed. Preliminary results from two experiments lend support to this notion.

In the first experiment subjects were given 4 digits (randomly selected) and asked to form a visual image of the digits in two rows of two each. On half of the trials, the digits were to be imaged as large subjectively as possible while still keeping all of the digits in mind simultaneously, and on the remaining half of the trials they were to be imaged at the smallest size possible while still remaining identifiable. Following construction of the appropriate images, a subject received one of two types of trials. In one case, a digit was presented, and he was to "see" if it was included in the imaged set. In the other case, in constrast, an animal-property pair (e.g., lion-feet) was presented. In this circumstance a subject was to forget about the digits and verify whether or not the animal had that property by referring to an image of the beast. If the digit was included in, or property appropriate for, the image, the subject was to respond "true" as quickly as possible (after consulting an image), otherwise he was to respond "false." The results for digit probes were as before: A probed digit included in a large image was detected more quickly than when the sought digit was represented in a small image. The data of real interest, however, were times to evaluate animal-property pairs following large or small images of digits. For "true" pairs, interestingly, more time was in fact required when subjects had constructed a large, compared to small, image of digits. This result is consistent with the hypothesis that images require time to fade.

The second experiment was identical to the first except that subjects did not vary the subject size of the imaged digits. Instead, the digits either were imaged as if they were bulky forms cut out of plywood, with very little space between them, or as if they were drawn with very narrow lines. All digits were to be imaged at the same size, however. Not surprisingly (given the results of the first experiment), animal-property pairs required more time to evaluate if the preceding image portrayed bulky plywood digits. Before any firm conclusions or inferences may be drawn from these data, however, appropriate controls must be conducted. When imagery is not required for evaluating an animal-property pair, for example, the imaginal size or form of preceding digits should be irrelevant. In addition to this sort of control, we presently are conducting a number of similar experiments and replications.

Concluding Remarks

A computer graphics metaphor of imagery seems to have some usefulness. Although this treatment is not really an "analogue" model, as such, it also is not simply a "propositional" model either. One might object to the necessity of an analogue surface representation, and attempt to model imagery in terms of networks of propositions or the like. Although something like our display must be postulated, it seems to me, in order to explain the phenomenology of experiencing an image, it need not necessarily be required to account for how imagery functions in human cognitive processes (see Baylor, 1971; Moran, 1973, Pylyshyn, 1973). It is the task of future experiments to justify inclusion of an analogue stage in image processing. Two sorts of research currently are in progress in my lab that bear on this issue.

The analogue supposition would receive support if we demonstrated that pictorial properties of imagery actually have some psychological consequences. Some results reported in Kosslyn & Alper (in preparation) seem to implicate the pictorial properties of imagery as an important factor in memory. In these experiments people imaged pairs of objects with both objects pictured either at normal, appropriate relative sizes or with the second object imaged tiny relative to the first. When subjects imaged in the latter way, their memory for the second object named in a pair (when given the name of the first object as a recall cue) was much poorer than when size was not distorted. Similarly, if the name of an object that was imaged tiny was used as a recall cue, recall of the first object in the pair (itself imaged at full size) also was impaired. This result was obtained even when we told people how the objects should be interacting in their images (thus, the size effect is not an artifact of different relations being incorporated in the two size conditions). If the surface pictorial properties of imagery are merely epiphenomenal, these results are difficult to explain. If images, once constructed, then may be encoded into memory, these results are not surprising given Kosslyn´s (in press) findings (i.e., smaller things are "harder to see"). The Kosslyn & Ralph results on image overflowing cited earlier also seem difficult to account for without reference to some sort of analogue display.

The second sort of data that seems to argue for a level of analogue display concerns the scanning of visual images. Earlier work (Kosslyn, 1973) demonstrated that if one must scan further on an image to locate a property, more time is required to verify the appropriateness of that property for the imaged object. Only three distances were used in this experiment, however, and simple network models seem capable of accounting for the data (in terms of distances between nodes in the net). The present work involves requiring scanning across 21 different interpoint distances between locations on an imaged map. If the actual distances predict time to shift from one point to another, a network model seems hard pressed to account for these results. Such a representation might postulate that dummy nodes exist between those representing the locations; longer distances would be separated in the net by more intervening nodes. Another experiment currently underway seems to preclude easy application of such a model to all instances of image scanning. This experiment involves scanning from the mouth to the eyebrows (and then categorizing them in various ways) on an imaged face. Four versions of each face are used, which vary in how low, how close to the mouth, the eyes and eyebrows are (the same sized head always is drawn, resulting in the appearance of individuals with more-or-less large foreheads in the different versions). No details of a face are changed from version to version other than actual metric distance between the mouth and eye regions. This variation would not seem likely to distort an underlying propositional representation of a face in a way that would lead one to predict shorter times to scan shorter distances on an image in less time. If these results prove reliable, imagery phenomena may not be as simply explained as some of us had hoped.

REFERENCES

Baylor, G.W., "A treatise on the mind´s eye." Unpublished Ph.D. thesis, Carnegie-Mellon University, 1971.

Collins, A.M., & Quillian, M.R., "Retrieval time from semantic memory." Journal of Verbal Learning and Verbal Behavior, 1969, Vol. 8, 240-247.

Conrad, C., "Cognitive economy in semantic memory." Journal of Experimental Psychology, 1972, Vol. 92, 149-154.

Kosslyn, S.M., "Scanning visual images: Some structural implications." Perception and Psychophysics, 1973, Vol. 14, No. 1, 90-94.

Kosslyn, S.M., "Constructing visual images: An exercise in neo-mentalism." Unpublished Ph.D. thesis, Stanford University, 1974.

Kosslyn, S.M., "Information representation in visual images." Cognitive Psychology, in press.

Moran, T.P., "The symbolic nature of visual imagery." Paper presented at the Third International Conference on Artificial Intelligence, Stanford University, August, 1973.

Pylyshyn, Z.W., "What the mind´s eye tells the mind´s brain: A critique of mental imagery." Psychological Bulletin, 1973, Vol. 80, No. 1, 1-24.

Smith, E.E., Shoben, E.J., Rips, L.J., "Structure and process in semantic memory: A feature model for semantic decisions." Psychological Review, 1974, Vol. 81, No. 3, 214-241.

Winograd, T., Understanding Natural Language. New York: Academic Press, 1973.

Winograd, T., "Frame representatons and the declarative/procedural controversy." In Bobrow and Collins (eds.) Representation and Understanding. New York: Academic Press, 1975.