Journal of Experimental Psychology
1973, Vol. 97, No. 1, 22-27

SEARCHING FOR OBJECTS IN REAL-WORLD SCENES[1]

IRVING BIEDERMAN,[2] ARNOLD L. GLASS,[3] AND E. WEBB STACY, JR.

State University of New York at Buffalo

The speed at which a single object can be detected in a real-world scene was reduced when the scene was jumbled compared to when it was coherent. Jumbling was most disruptive when the target object was not in the scene but had a high probability of occurring in that kind of scene. These results are discussed in terms of the possible role played by schemas in the processing of information from real-world scenes.

This experiment was designed to provide some characterization of how people search for objects in real-world scenes. The rapidly growing literature on visual search has been concerned almost exclusively with the speed (RT) and accuracy of processing displays of unrelated items. Models derived from the results of this research typically conceptualize an independent processing of the various positions of the display (e.g., Egeth, Jonides, & Wall, 1972; Rumelhart, 1970). Biederman (1972), however, has recently demonstrated that the overall coherency of an object's setting affected the recognition accuracy of that object. In that experiment, *S*s briefly viewed pictures of many varied scenes. Their task was to identify which object occupied a given cued position in the scene. Immediately after the presentation of the scene, an arrow was presented which pointed to a position where an object had been. The *S*'s task was to indicate, by pointing to one of four object pictures, which object had been cued. All four objects were from the scene just viewed. In that experiment, coherency was manipulated through a scene-jumbling procedure (described below), which destroyed the natural spatial relations of the entities of the scene. Jumbling was shown to reduce the accuracy of identifying the cued object even when *S* knew what objects to look for (in a condition where the four response alternatives, the pictures of objects, were provided prior to the presentation of the scene) and, to a lesser extent, where to look (in a condition where the arrow preceded the scene). Thus, jumbling was most likely affecting perceptual recognition and not just memory or response selection.

The present experiment explored the effect of this jumbling variable on a speeded search task. The experiment was also designed to provide *S* with some opportunity to use his overall characterization of the scene to mediate his detection performance. A secondary purpose of the experiment was to provide some estimate as to how long it takes to find a well-defined object in a scene.

METHOD

Subjects. The *S*s were 36 students from the State University of New York at Buffalo. Participation was part of their introductory psychology course requirement.

Scenes. The scenes were projected from 35-mm. black and white slides. A wide variety of scenes was sampled, e.g., streets, kitchens, desk tops, store counters, etc. Coherent and jumbled versions of each scene were made by photographing a 20 × 35 cm. print, which had been cut into six sections (generally with one horizontal and two vertical cuts) in such a manner as to leave at least four well-defined objects intact (Figures 1 and 2). The

[1] The authors would like to thank Sharon L. Cook for her invaluable assistance in the preparation of the stimuli and data collection, James Meeker for his help in running subjects, and Edward E. Smith for his excellent comments on an earlier draft of this paper. This research was supported by Grant 050-7201-A from the Research Foundation of the State University of New York and a National Institute of Mental Health Special Fellowship MH-50632-01 to the first author for a leave to Stanford University. The first author would like to express his appreciation to the Department of Psychology at Standord University for their stimulation and support during the writing of this paper.

[2] Requests for reprints should be sent to Irving Biederman, Department of Psychology, State University of New York at Buffalo, 4230 Ridge Lea Road, Buffalo, New York 14226.

[3] Now at Stanford University.

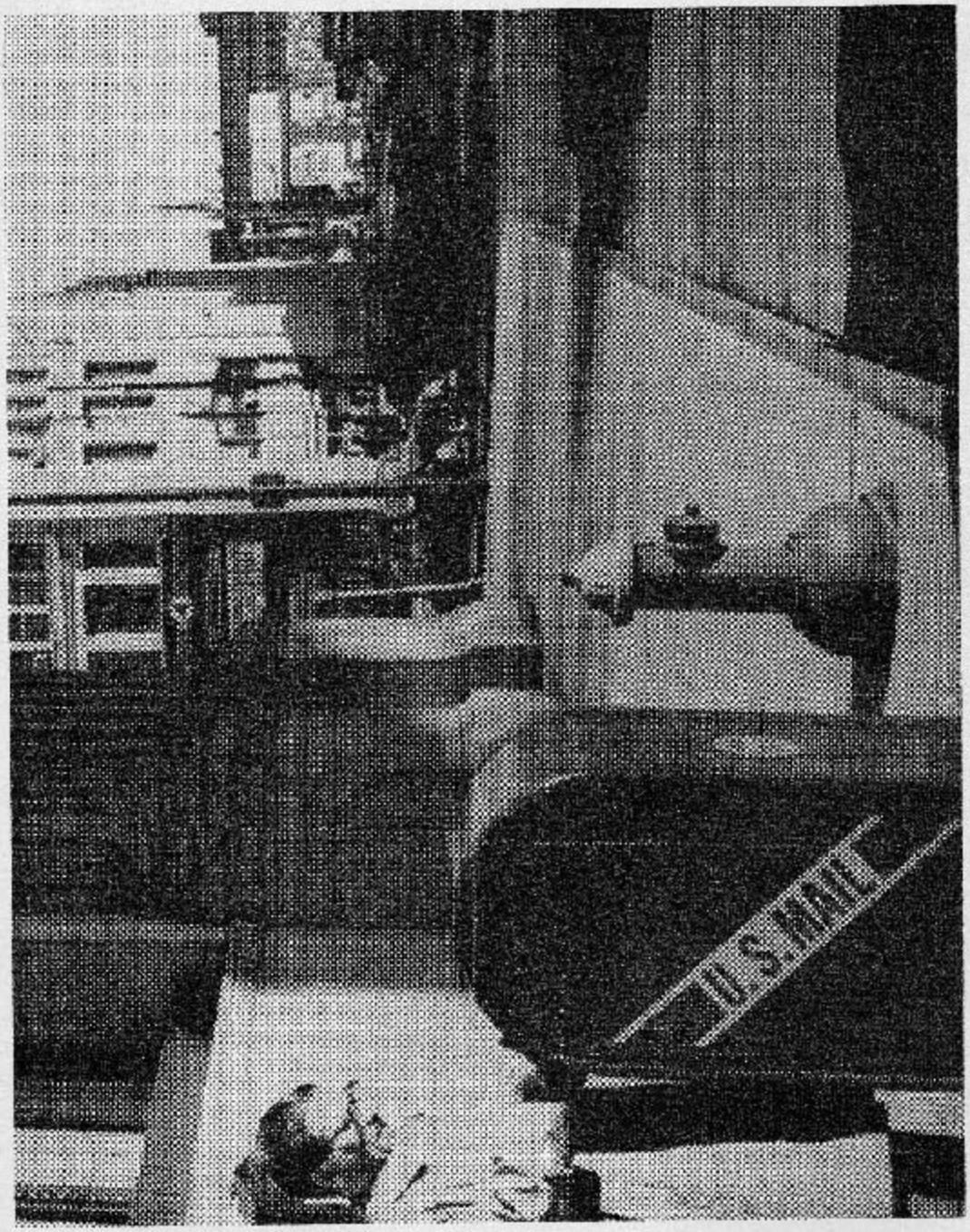

FIG. 1. Sample of coherent scene. (Note that the target object, a relatively easy one in this example, was the fire hydrant.)

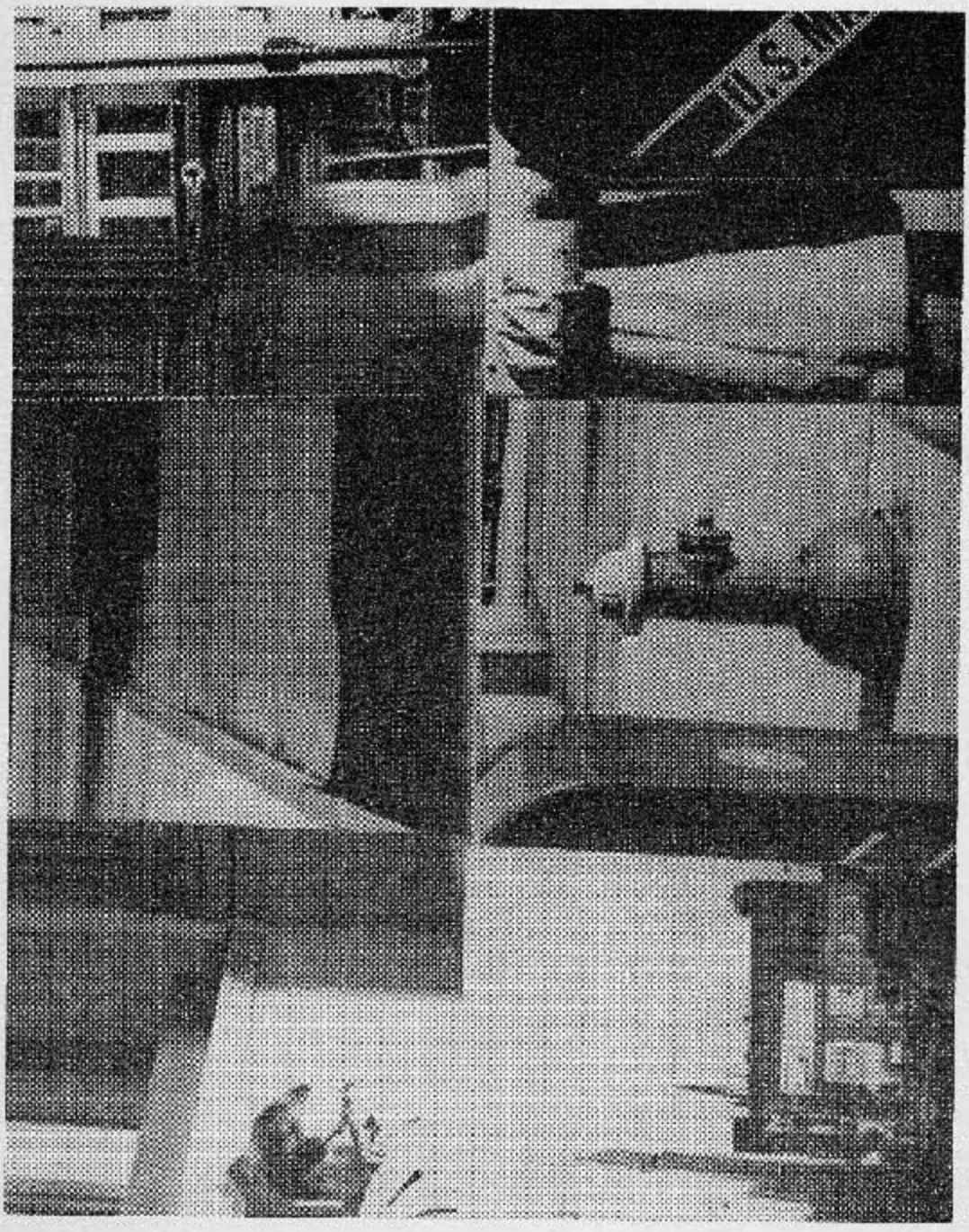

FIG. 2. Sample of jumbled scene. (Note that the lower middle section, which contains the target object, is the same as in Figure 1.)

coherent slide was taken after the sectioning of the print so that the section lines appeared in both versions. The sections of the jumbled version were rearranged (but never rotated) so as to destroy the natural spatial relations of the components. However, in the jumbled scene, one section was always left in its original position. This section always contained at least one well-defined object. The position of the section remaining constant was balanced across the different sections so that, for example, for one sixth of the scenes the top left section was identical for both the jumbled and coherent versions.

Design and procedure. Each *S* viewed 112 scenes, half jumbled and half coherent, in random-appearing order. The first 16 scenes were considered as practice trials and were not included in the data analyses. The remaining 96 scenes were identical to the scenes used in the Biederman (1972) experiment. The scenes subtended a visual angle of 19° as projected by a Kodak Carousel projector fitted with a fast rise-time shutter. Before viewing each scene, *S*s perused (for approximately 5 sec.) a card with the picture of an object on it. This object was cut out of the original photographs used in making the slides. Upon presentation of the scene, *S* pressed either a "yes" or "no" finger key as soon as he judged whether the object was or was not in that scene. On one third of the trials, the object was, in fact, in the scene ("yes" responses). On these trials, the target object was from the section which remained in its original position (and was the cued object in the Biederman, 1972, experiment). On another third of the trials, the object was not in the scene but could have been in it ("possible-no" responses). For example, the target object might have been an automobile and a street scene might have been projected, but the automobile was not one of the automobiles in the scene. Or the object might have been a cup and the scene was that of a kitchen, but the cup was not in the scene. On the remaining third of the trials ("impossible-no" responses), the object was one that was highly unlikely to have appeared in the scene. (In fact, the object never did appear in the scene.) Examples of impossible-no trials would be a cup as a target in a street scene, or an automobile as a target in a kitchen scene. For a given *S*, the six types of trials (two kinds of scenes and three kinds of response categories) were presented in random appearing order. Each *S* viewed only one of the two versions of a given scene (jumbled or coherent) and responded to that version under only one of the three response categories. The targets were balanced across *S*s so that each target was used in the six conditions an equal number of times. The trials were self-paced; the scenes remained projected until a response was made or 8 sec. elapsed.

Results

For each *S*, median correct reaction times (RTs) were calculated for each of the six kinds of trials. The means of these

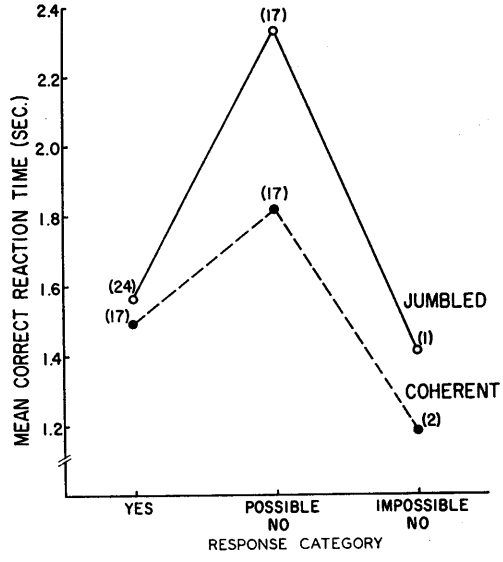

FIG. 3. Mean correct reaction times as a function of scene version and response category. (Numbers in parentheses are percent errors.)

medians, as well as the mean percent errors, are shown in Figure 3. Approximately 1.5 sec. are required to find an object that is in a scene. The main effects on RTs of coherency, F (1, 35) = 12.22, $p < .01$, and response category, F (2, 70) = 34.46, $p < .001$, were both significant, as was their interaction, F (2, 70) = 3.68, $p < .05$. Quite striking was the effect of probability in the "no" response categories: The impossible-no responses were, on the average, about .75 sec. faster than the possible-no responses.

The coherency variable was also examined by collapsing across *S*s and calculating a mean RT for the jumbled and coherent version of each slide. Of the 96 experimental slides, 71 had shorter RTs with the coherent versions and 25 had shorter RTs with their jumbled versions, $\chi^2(1) = 22.04$, $p < .001$.

Errors were almost nonexistent in the impossible-no response category but very high in the yes and possible-no categories. A significant Coherency × Response Category interaction, F (2, 70) = 8.05, $p < .001$, in the error rates is completely attributable to the higher error rates in the jumbled yes responses compared to the

coherent yes responses. If a speed for accuracy trade off were operating for the yes responses, then the effect of coherency on this response category was underestimated.

Discussion

A number of recent theories of stimulus recognition (e.g., Eden, 1962; Halle & Stevens, 1964; Neisser, 1967) postulate a multistage theory in which an initial holistic characterization of the stimulus (determined by gross feature tests and context) biases the subsequent testing, weighting, and combination of detailed features and the subsequent memory representation. The idea that a holistic representation can precede—and facilitate—the processing of specific parts is a fundamental tenet of analysis by synthesis models of perceptual recognition. One set of speculations to account for the results of both the present and Biederman (1972) experiments that is consistent with such a multistage theory would emphasize the importance of the finding that the possible-no responses were more affected by jumbling than the impossible-no responses (504 vs. 227 msec.). In the impossible-no condition, as soon as *S* could achieve an overall characterization (or schema) of the scene—as soon as he could recognize the scene as a street, desk top, or kitchen—he had sufficient information to respond. That *S* did tend to exit at this point is evidenced by the short impossible-no RTs (relative to the possible-no RTs). The 227-msec. effect of jumbling for these responses could be attributed to a delay in the attainment of this overall characterization.

That jumbling might also have slowed the recognition of specific objects is consistent with the additional effect of jumbling on the possible-no RTs. Here achieving a schema was insufficient and *S* would have to engage in detailed feature processing and object identification to determine if the target was in the scene. There are a number of nonexclusive ways in which the additional effect of jumbling could have occurred. Jumbling could have interfered directly with feature processing and object identification. Or the effect could have been an indirect one, if feature processing and object identification were sometimes mediated by a schema (Bruner & Potter, 1964; Bugelski & Alampy, 1961).[4] In this latter case, jumbling could have slowed the formation of a schema or yielded a schema that was of less value for object identification.

That a schema can affect the processing of a scene is an old notion (Bartlett, 1932) currently enjoying a revival. One reason for this is that many workers in artificial intelligence have come to the realization that if their machines are to be capable of understanding a scene or sentence, then the machine must have some internal model with which to interpret ambiguous inputs (Michie, 1971; Winograd, 1972). A program such as Winograd's achieves its success largely through its capacity to apply knowledge of its world to disambiguate inputs specified in terms of more primitive features.

Until the nature of jumbling is resolved, the above account of the results of this experiment remains highly speculative. Consistent with the results are explanations that do not require any effect of a schema on perceptual recognition. Two of the more obvious and tractable ones will be briefly considered:

1. There are a greater number of object fragments in a jumbled scene compared to its coherent version. While the fragments did not overlap with the target object, they could have "drained" some of the processing capacity. This interpretation can be tested through a procedure whereby scenes are drawn or cropped in such a manner that jumbling leaves the number of entities intact.

2. Jumbling might have disrupted an informative external scanning of the scene. For example, if the target was a cup in a kitchen scene, one would tend to look at the counter tops rather than on the floor. However, it is not likely that all of the effect of jumbling can be attributed to differential eye movements since I. Biederman, E. W. Stacy Jr., and A. L. Glass (unpublished manuscript, 1972) found effects of jumbling at exposure durations (20 msec.) far too brief for an eye movement. Nonetheless, the present experiment would have been more adequate if the scenes had been presented for only a brief duration. Moreover, a position could have been cued (as in the Biederman, 1972, experiment) and *S* could have responded "yes" or "no" depending on whether the cued object was the target.

These two explanations are consistent with a theory positing an independent perceptual

[4] If the identification of objects in scenes is sometimes mediated by a schema of that scene, then it might be expected that the presence of an inappropriate schema might *lengthen* RTs. Such an effect could, perhaps, be detected with an "impossible-yes" category in which unlikely objects were targets, e.g., a chandelier in a barn.

processing of the various positions of a scene within each fixation. Future research will be necessary to resolve the issue as to the role of higher order units in the perceptual recognition of scenes.

REFERENCES

BARTLETT, F. C. *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press, 1932.

BIEDERMAN, I. Perceiving real-world scenes. *Science*, 1972, **177**, 77–80.

BRUNER, J. S., & POTTER, M. C. Interference in visual recognition. *Science*, 1964, **144**, 424–425.

BUGELSKI, B. R., & ALAMPAY, D. A. The role of frequency in developing perceptual sets. *Canadian Journal of Psychology*, 1961, **15**, 205–211.

EDEN, M. Handwriting and pattern recognition. *I. R. E. Transactions on Information Theory*, 1962, IT-**8**, 160–166.

EGETH, H., JONIDES, J., & WALL, S. Parallel processing of multielement displays. *Cognitive Psychology*, 1972, **3**, 674–698.

HALLE, M., & STEVENS, K. N. Speech recognition: A model and a program for research. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language: Readings in the philosophy of language.* Englewood Cliffs, New Jersey: Prentice-Hall, 1964.

MICHIE, D. *On not seeing things.* (Experimental Programming Report No. 22) Edinburgh, Scotland: University of Edinburgh, Department of Machine Intelligence and Perception, 1971.

NEISSER, U. *Cognitive psychology.* New York: Appleton-Century-Crofts, 1967.

RUMELHART, D. E. A multicomponent theory of the perception of briefly exposed visual displays. *Journal of Mathematical Psychology*, 1970, **7**, 191–218.

WINOGRAD, T. Understanding natural language. *Cognitive Psychology*, 1972, **3**, 1–191.

(Received March 31, 1972)