Journal of Experimental Psychology
1974, Vol. 103, No. 3, 597–600

ON THE INFORMATION EXTRACTED FROM A GLANCE AT A SCENE[1]

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

State University of New York at Buffalo

Pictures of jumbled and coherent versions of real-world scenes were shown for durations of 20–300 msec. In Experiment I, *S*s selected that label from a pair of labels which they judged to better describe the scene. The more similar the label pairs, the less accurate were the choices. In Experiment II, a cue was presented immediately after the scene which designated one of the objects in the scene. Jumbling reduced both the accuracy of identification of the cued objects and the accuracy of descriptor choice in Experiment I. The results are discussed in terms of two kinds of information that *O* extracts from a single fixation of a scene: (*a*) individual objects and (*b*) an overall characterization of the scene.

What kinds of information can be extracted from a glance—a single fixation—at a scene? The present investigation was concerned with two possible approaches to this question: (*a*) a measure of the accuracy with which an observer is able to select a verbal label that describes the "topic" of the scene, and (*b*) a measure of the accuracy of the identification of a sample—an object—from a scene. The former method was used by Biederman (1972), who compared the accuracy of object identification in coherent and jumbled scenes. Also, Biederman, Glass, and Stacy (1973) measured the amount of time required to detect a single object in a scene. However, in both experiments, the scenes were presented for durations sufficiently long for eye movements to play a role. The present experiments were designed to study scene perception at presentation durations too brief for an eye movement.

Biederman et al. (1973) reported that searching for an object in a coherent scene required less time than searching for the same object in a jumbled version of that scene. Evidence that jumbling reduced *S*'s ability to derive a topic or overall conceptualization of the scene came from an examination of the effect of jumbling on those trials when the object was, in fact, not in the scene. Both for jumbled and coherent versions of a scene, *S*s were considerably faster—by about 750 msec.—when the target object was very unlikely to appear in the scene (e.g., when a cup was the target in a street scene) as compared to when the target was likely to appear in the scene (e.g., a cup in the kitchen). Biederman et al., reasoned that when viewing both jumbled and coherent scenes *S*s were able to derive an overall conceptualization of the scene and use it to quickly terminate their searches when the objects were unlikely, given the conceptualization. That jumbling affected the speed at which *S*s were able to achieve that conceptualization was evidenced by a relatively large 225-msec. effect of jumbling on these "impossible–no" reaction times (RTs).

The jumbling variation employed in the Biederman (1972) and Biederman et al. (1973) experiments was also incorporated into the design of the present investigation. While the effects of jumbling on the processes by which information is extracted from a scene are yet to be specified, the inclusion of this variable is predicated on the belief that the perception of a scene might be something more than the perception of the sum of the individual objects that comprise the scene. An obvious analogy can be made with the words of a sentence or paragraph. Just as the sequence of the words in a sentence or paragraph or utterance is critical to the reader's or listener's developing an overall conceptualization of the passage, the spatial relations among the various entities of a scene are critical for the development of a conceptualization or schema of a scene.

EXPERIMENT I

Experiment I in this investigation was designed to obtain a more direct assessment of the effect of

[1] The research was supported by Grant 050-7201-A from the Research Foundation of the State University of New York. The authors thank Edward E. Smith for his helpful comments and Sharon L. Cook for her help in running *S*s and analyzing data.

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

[3] Now at Stanford University

jumbling on the confirmation of an overall conceptualization of the scene. Specifically, the accuracy of selecting a member from a pair of verbal descriptors of the scene was studied. The similarity of these labels was varied in an attempt to affect the diagnostic value of individual objects. When the pairs of labels were similar (e.g., "shopping plaza" vs. "busy road and stores") the individual objects were of less informational value than when the pairs were dissimilar (e.g., "shopping plaza" vs. "kitchen"). If holistic characteristics of a scene are required to discriminate among similar labels, jumbling should be more disruptive with such labels than when individual objects are of high diagnostic value, i.e., when labels are dissimilar.

Method. The stimulus materials, 35-mm. positive slides of a wide variety of scenes, were described in Biederman (1972) and Biederman et al. (1973). For each scene there was a coherent and jumbled version—the jumbled version was constructed by sectioning the original into six parts and repositioning (but never rotating) five of them.

Each of 32 *S*s viewed a sequence of 112 jumbled and coherent scenes. The first 16 were considered practice trials and were not included in the data analysis. The scenes were projected, with a visual angle of 10°, onto a screen by a projector fitted with a tachistoscopic shutter. Immediately after the presentation of the scene a masking grid was presented for 300 msec. Following each presentation, *S* selected the member from a pair of verbal labels (words or phrases) which he judged to be the more accurate description of the scene. The 116 pairs of labels were numbered and printed in a booklet which *S* perused before each presentation. Half of the label pairs had phrases that were dissimilar, such that it would be unlikely for an object appropriate for one description to be present in the other. Examples were, "lawn in back of house" vs. "kitchen," "vanity table" vs. "mountains," "fireplace" vs. "parking lot in front of stores." The other half of the labels were similar, such that objects in one scene might be reasonably expected to be in the other. Examples are, "shopping plaza" vs. "busy road and stores," "bedroom" vs. "livingroom." Different sequences of scenes and descriptions were used so that scene version (jumbled or coherent), label similarity (similar or dissimilar), and presentation duration (20, 50, 100, or 300 msec.) were all balanced over the 96 scenes. In addition to the 32 *S*s who viewed the scenes at the 20–300 msec. durations, a panel of eight judges viewed the scenes at a presentation duration of 4 sec.

Results. Figure 1 shows the mean percentage of correct choices as a function of scene version, duration, and label similarity. All three variables had highly reliable main effects: jumbling, F (1, 31) = 19.75, $p < .001$, duration, F (3, 93) = 32.55, $p < .001$, and label similarity, F (1, 31) = 83.95, $p < .001$. (The analysis of variance did not include the 4-sec. data.) The only interaction that was significant was Duration × Label Similarity, F (3, 93) = 5.09, $p < .01$. The Jumbling × Duration interaction fell just short of significance, F (3, 93) = 2.74, $.05 < p < .10$. In both cases a relatively small effect of jumbling and similarity at the shorter durations grew to a larger one at the longer durations. The effects of jumbling and label similarity would have been still larger but for a ceiling effect at the longer durations. Thus, at the 300-msec. duration, 22 (of 32) *S*s were at 100% accuracy in the coherent–dissimilar condition, 13 *S*s in the jumbled–dissimilar condition, 14 in the coherent–similar condition, and only 1 in the jumbled–similar condition.

At the 300-msec. presentation duration, jumbling was particularly disruptive for the similar label condition. Moreover, there was no improvement in accuracy in the jumbled–similar condition between 50 and 300 msec. The significant Jumbling × Label Similarity interaction at 300 msec., F (1, 31) = 8.87,

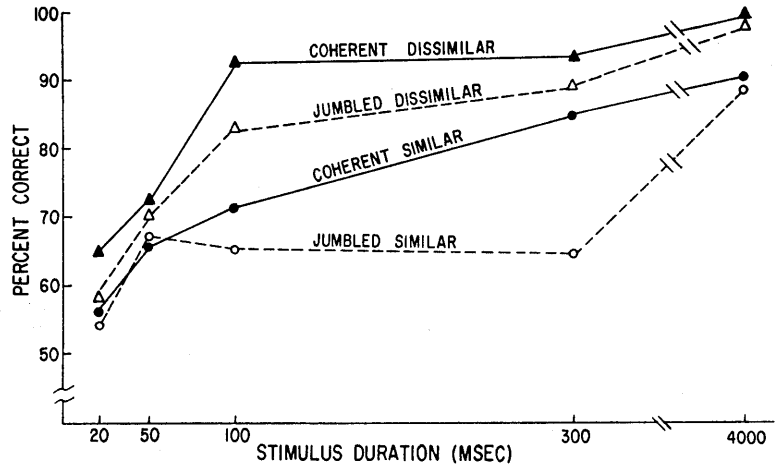

FIGURE 1. Mean percentage of correct choices of a scene label as a function of scene version (jumbled vs. coherent), label similarity, and presentation duration.

$p < .01$, revealed that the similar labels were more affected by jumbling than the dissimilar labels. However, the ceiling effects discussed previously may have contributed to this interaction, so caution must be exercised in interpreting it.

Experiment II

Selecting a label to describe a scene is, in part, likely to be dependent upon the identification of individual objects in a scene. Experiment II was designed to describe the time course of object identification over the same range of durations used in Experiment I. This experiment thus extends the paradigm employed in the Biederman (1972) experiment to a range of durations too brief for an eye movement. In addition to the picture choices, a condition was included in which *S* responded by selecting among verbal labels of the objects. This variation was designed to assess the degree to which *S* relied on simple physical features (size, brightness, texture, etc.) to make his response.

Method. The scenes and presentation equipment were the same as those in Experiment I. The section which was not moved contained at least one well-defined, easily nameable object. This object was the one designated by the cue. Since it was in the section which remained constant, the cued object occupied the same position in both the coherent and jumbled versions.

The procedure and design closely followed that of the Biederman (1972) experiment. Immediately after the presentation of a scene a combined masking-grid and cue (circle) slide was presented for 300 msec. The circle occupied the same position as the object in the scene from the section which was not moved. The *S* was to indicate which object occupied the position designated by the cue. Half of the *S*s, as in the Biederman (1972) experiment, responded by designating one of four object pictures, mounted on a single page in a photo album. The other *S*s responded with one of four object names printed on a card. Since *S*s were allowed to peruse the response alternatives prior to the presentation of the scene, this experimental procedure is analogous to the alternatives–before and cue–after condition of the Biederman (1972) experiment.

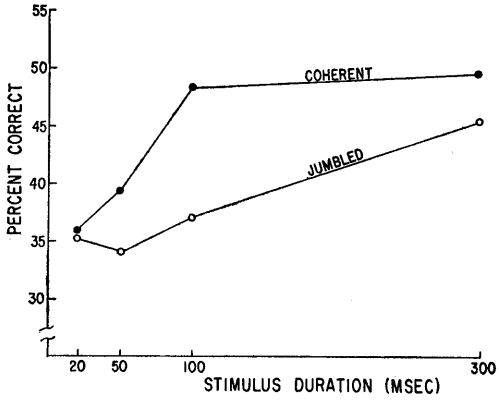

FIGURE 2. Mean percentage of correct object identifications as a function of scene version (jumbled vs. coherent) and presentation duration.

Results. Figure 2 shows the mean percentage of correct identification as a function of scene version and duration. The effects of both scene version and duration were highly significant, F (1, 62) = 18.35, $p < .001$, and F (3, 186) = 19.92, $p < .001$, respectively. Picture alternatives led only to slight (41.4% vs. 39.8%) and nonsignificantly F (1, 62) < 1.00, higher accuracy than labels. It is likely, then, that *S*s were fully identifying the objects and not responding on the basis of simple physical features. The only interaction that was significant was Scene Version × Duration, F (3, 186) = 2.80, $p < .05$; the advantage of coherent over jumbled scenes increased from the 20-msec. to 100-msec. presentation duration, but coherent performance then leveled off, while the jumbled scenes showed some increase in accuracy. The data for the picture alternatives at 300 msec. were almost identical to the corresponding condition in the Biederman (1972) experiment. In that experiment, the accuracy neither consistently nor reliably improved as presentation duration was increased from 300 to 700 msec. This lack of improvement at the longer durations most probably represents a ceiling effect imposed by the inclusion of scenes where the cued objects were small and peripheral.

Discussion

From an exposure duration of 100 msec.—from a single fixation on a scene—*O* is able to identify 45%–50% of the moderately sized objects. The information from that fixation is also sufficient to select correctly a descriptor of the topic of the scene from a dissimilar alternative on 95% of the presentations and from similar alternatives on 70% of the trials. Jumbling reduced the accuracy of performance on both of these tasks.

The investigation was not designed to provide a theoretical accounting of scene perception but nevertheless the data can be described in terms of two kinds of information which might have been available to *S* in performing the label-matching task (Experiment I).

First, *S* might have simply sampled objects and inferred the correct answer on the basis of the objects. Object sampling would have been most beneficial for the dissimilar label pairs, since in these cases almost any of the objects from the scene would be diagnostic toward inferring the correct label. The effect of jumbling in the dissimilar alternatives condition likely reflects the effect shown in Experiment II where jumbling reduced the accuracy of object identification. However, with similar response alternatives, object sampling would be less valuable. Perhaps *S* also employed a second, more holistic mode of processing which made use of the spatial relations among the objects in the scene to help him distinguish, say, "shopping plaza" from "busy road

and stores." Jumbling should have a particularly disruptive effect on a condition where *S* must rely on a holistic mode of processing. From this perspective, it is interesting to note that while there was a considerable gain in accuracy of identifying objects in jumbled scenes (Experiment II) between 50 and 300 msec., there was no improvement in accuracy in the jumbled–similar condition between 50 and 300 msec. (Experiment I). Moreover, at 300 msec., the effect of jumbling was greater for the similar descriptors than for the dissimilar descriptors. But as mentioned previously, ceiling effects may have contributed to this interaction.

If there is any truth to the above descriptions of a dual mode of processing, then it is likely that *S* simultaneously handles the information from a scene with both modes. That is, individual objects would be identified along with the attainment of the overall scene characterization. At this time it is not clear how such a characterization is achieved. Its development might be aided by having a consistent array in depth along with identification of some objects and their relations to other objects.

REFERENCES

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

Biederman, I., Glass, A. L., & Stacy, E. W., Jr. Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 1973, **97**, 22–27.

(Received March 6, 1974)