Perception, 1996, volume 25, pages 343–352

Spatial context in recognition

Moshe Bar¶, Shimon Ullman
Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: bar@selforg.usc.edu; shimon@wisdom.weizmann.ac.il
Received 1 May 1995, in revised form 4 December 1995

Abstract. In recognizing objects and scenes, partial recognition of objects or their parts can be used to guide the recognition of other objects. Here, the role of individual objects in the recognition of complete figures and the influence of contextual information on the identification of ambiguous objects were investigated. Configurations of objects that were placed in either proper or improper spatial relations were used, and response times and error rates in a recognition task were measured. Two main results were obtained. First, proper spatial relations among the objects of a scene decrease response times and error rates in the recognition of individual objects. Second, the presence of objects that have a unique interpretation improves the identification of ambigous objects in the scene. Ambiguous objects were recognized faster and with fewer errors in the presence of clearly recognized objects compared with the same objects in isolation or in improper spatial relations. The implications of these findings for the organization of recognition memory are discussed.

1 Introduction

Natural scenes usually contain multiple objects and different scenes (such as a street, an office, etc) are associated with different groups of characteristic objects. The information about typical arrangements, typical members in a scene, and the typical spatial relations among them may be used during the process of object recognition and scene interpretation. Our goal in the present study was twofold. First, we wanted to examine the effect of the spatial relations between two objects that often co-occur in the same scene on performance in recognition tasks. In particular, we wanted to examine whether facilitation in recognition depends solely on the identity of the objects, namely, the presence of one object facilitates the recognition of related objects, or whether such effects depend on the spatial relations between the objects in question. This question has important ramifications for the organization of recognition memory. Second, we wanted to examine whether an object that has a unique interpretation can disambiguate the identity of a more ambiguous object, and how it can influence the recognition of the complete configuration.

Questions related to the effect of spatial relations on recognition performance were examined in the past (Biederman 1972, 1981; Biederman et al 1982; Cave and Kosslyn 1993; Hock et al 1978; Mandler and Johnson 1976; Palmer 1975a). Two main differences between these and the current study are noteworthy. First, researchers in previous studies were mainly concerned with visual search and detection tasks, whereas in the present study we extend the question to recognition tasks. The main difference between search and recognition tasks is that in search the object to be found is known a priori, only its location must be determined. In recognition, identity of the object must be determined. When trying to recognize an object, we are performing a model-selection process; we must somehow 'scan' a large number of existing models, as opposed to in the search task, where we have a clear idea as to what we are looking for.

¶ Present address: Department of Psychology, University of Southern California, Hedco Neurosciences Building, Los Angeles, CA 90089-2520, USA

Thus, it is reasonable to assume that the cognitive mechanisms involved in recognition tasks may be different from those subserving the search tasks. Second, the setting we used is also unique in the sense that instead of using a single object (as in Cave and Kosslyn 1993) or complete scenes (as in Biederman 1972, 1981; Mandler and Johnson 1976; Palmer 1975a), our configurations consisted of two separate objects. This allowed us to study more systematically the interactions between individual objects, and to attribute the observed effects to the interactions between objects, rather than other possible scene-configuration effects. For example, in Biederman (1972) real-world photographs were cut and jumbled in order to investigate the effect of spatial relations among objects in a scene. However, in addition to the spatial relations and the context, the original real-world scene is rich in information regarding shapes, shading, textures, etc. Consequently, in addition to the jumbling of the spatial relations, abnormal discontinuities in scene properties may arise, and it is difficult to assess their effect on the overall performance.

Experimental evidence regarding the facilitation of recognition by the use of typical spatial relations and context may also improve artificial recognition methods. Computer vision systems typically involve a set of independent object models stored in memory in an unstructured manner, and recognition is performed by comparing each image object with each of the models in memory (Grimson 1990; Lowe 1986; Ullman 1989). As we will suggest in section 3.2, the use of typical spatial relations and context information could significantly reduce the search for the appropriate object model, could facilitate recognition in multiobject configurations, and help coping with degraded, ambiguous, or missing information in the image. With these differences and potential benefits in mind, the current experiments were designed to investigate the effect of recognizing one object on the recognition of related objects, and to examine the dependence of such an effect on the spatial arrangement. We also tested the influence of an object with a clear unambiguous interpretation, a 'key object', on the recognition of a related object nearby.

2 Experiment 1

2.1 *Method*

2.1.1 *Subjects.* Eighteen graduate students participated as volunteers. All had normal or corrected-to-normal vision, and none was aware of the purpose or predictions of the experiment.

2.1.2 *Materials.* Sixty-three stimuli were derived from the eight figures depicted in figure 1. Three types of stimuli were created: single objects, two objects that maintained their proper spatial relation (proper relation), and two objects with an improper spatial relation (improper relation) (figure 2). By 'objects' we mean here objects that are parts of the figures, such as a hat, a hand, an envelope, etc. The two objects that participated together in either proper-relation or improper-relation displays were selected in such a way that one of them was easier to identify (the 'key object'), while the second was more difficult to identify (the 'ambiguous object').

All objects used in the experiment were first tested in a pilot study with ten subjects by means of a free-viewing spontaneous-naming task. Ambiguous objects were defined as objects that either had several possible interpretations or had no clear interpretation at all, when presented in isolation. Key objects were defined as the objects for which there was a unique consistent interpretation. As we shall see, our initial selection of key and ambiguous objects was later supported by means of response times and error rates. The spatial relations were scrambled by placing the objects in random locations, keeping approximately the original physical distances.

The order in which the stimuli appear may affect the results; for example, the performances of a subject on recognizing improper-relation stimuli may differ when they appear before or after the proper-relation arrangement of the same objects. Therefore, we ordered the stimuli so that the proper-relation, improper-relation, and single-object stimuli of a certain figure were presented in different orderings for different subjects. Sixty-three stimuli, which consisted of thirty-three single-object, fifteen proper-relation, and fifteen improper-relation stimuli, were balanced in six different ways (all six possible combinations of ordering single-object, proper-relation, and improper-relation stimuli). The subjects saw the complete figures (in figure 1) only at the end of the experiment.

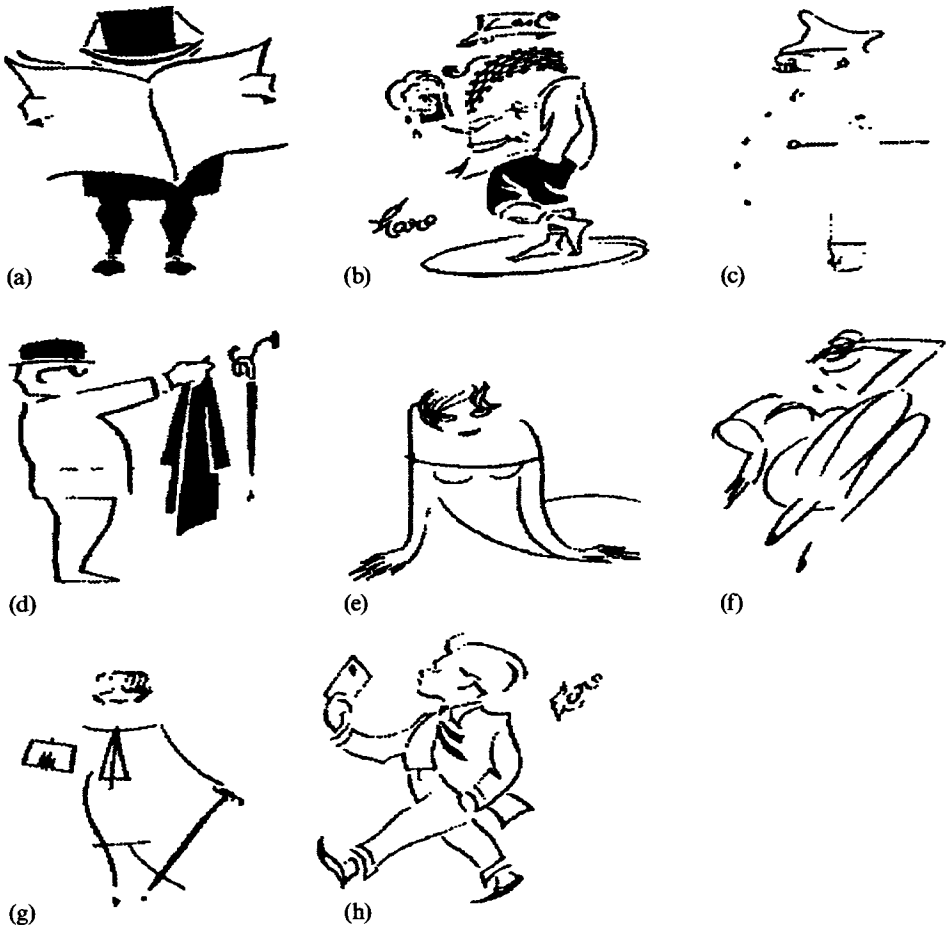

Figure 1. Examples of the original figures from which we produced our stimuli (reproduced with permission from Green and Courtis 1966). The objects that were selected as stimuli have a unique interpretation within the context of the complete figure.

2.1.3 *Apparatus.* The drawings were scanned by a MICROTEK scanner and manipulated by a Macintosh IICi computer. A Silicon Graphics Personal Iris 4D/35 computer, with a 1280 pixel × 1024 pixel resolution screen, controlled the stimulus presentation, and recorded response times and error rates.

Figure 2. Examples of the different stimuli. (a) Key objects (a leg on the left and a head on the right; (b) ambiguous objects (legs on the left and a glass of beer on the right); (c) proper spatial relations (hair with a hand holding an envelope on the left and a hat and a leg on the right); and (d) improper spatial relation (a hat and a leg on the left and a leg with glasses on the right). Subjects had to identify all objects in each stimulus.

2.1.4 *Procedure.* Prior to testing, subjects were given verbal instructions. The task was self-paced in that subjects began each trial by pressing a mouse button. The stimulus then appeared on the screen. Each stimulus remained visible until the subject responded (pressed a key). The computer recorded the time from the onset of the stimulus presentation to the beginning of the response. The descriptions given after the responses were recorded by the experimenter.

Subjects had to recognize all parts of the display and respond as quickly and as accurately as possible, by pressing a key, and then naming aloud the parts. They were instructed to respond only after recognition of the complete setting, or when they could not give any interpretation to the stimuli. The subjects were asked whether they

had any questions and all questions about the procedure were answered. Each subject viewed one of the six sets of sixty-three stimuli. Stimuli presentations were separated by a 3 s interval. The subjects sat approximately 40 cm in front of the computer screen. The stimuli subtended approximately 9.5 cm × 6.5 cm (6.77 deg × 4.65 deg of visual angle) on the screen. Theoretically, subjects could respond by pressing the key and continue the analysis after the image had disappeared. This possibility was reduced to minimum since subjects answered immediately after they pressed the key, and their instructions were to press the key only when they had recognized all the parts of the stimulus. In addition, the problem is the same in all four configurations and assumed to have a similar effect.

2.2 *Results*

Correct responses from all trials and response times from trials in which the correct names were produced were analyzed. Mean response times were calculated for each stimulus. Outliers were removed prior to analysis; an outlier was defined as response time that was greater by a factor of 3 than the mean of that condition without the outlier (we had only one subject whose responses had to be omitted). We also calculated the number of errors made in each condition. A response was considered erroneous when the subject was unable to produce any interpretation, or when it was different from the unique interpretation given to the same object in the context of the complete original figure.

For examples of correct responses consider the glass of beer depicted in figure 2b. Responses such as "ice cream" or "someone crying" were considered to be incorrect, whereas "glass" or "glass of beer" were considered as correct responses. Responses which described only part of the stimulus were also considered incorrect, eg "a leg" for figure 2d. Different descriptions such as "a man holding an envelope" and "hair, hand and envelope" for figure 2c were both correct. It is worth noting that the subjects' responses were highly consistent in the naming specificity they produced for a given image.

The definition of 'correct' is somewhat arbitrary but, as we shall see, it is sufficient for the purpose of our analysis. The mean response time for correct responses was 2.36 s, and the mean error rate was 26%.

2.2.1 *Response times.* In order to test the effect of spatial relations on the response time, we first considered the responses of subjects who correctly identified both the proper-relation and its corresponding improper-relation version (forty-eight cases). The comparison between the performance in the two versions is depicted in figure 3a. A paired (differences) t-test was performed on these data. The difference in mean response times was significant. Changing the correct spatial relations greatly increased naming times: objects that were presented in the improper spatial relations were correctly identified in 3.535 s on the average; objects that were presented in the proper spatial relations were correctly identified in 2.565 s on the average. The mean difference was 0.97 s ($t_{47} = -2.710$, $p < 0.009$).

We next considered the responses of subjects who responded correctly to the proper-relation and improper-relation versions, and also correctly identified the two objects presented separately (thirty cases). The comparison between performances in these four versions is depicted in figure 3b. On the average, the set of proper-relation stimuli was correctly identified in 2.183 s, the set of improper-relation stimuli in 3.577 s, key objects in 1.357 s, and ambiguous objects in 3.051 s. An analysis of variance (ANOVA) was performed on these data and it revealed a significant difference between the response times for correctly identifying proper-relation stimuli, key objects, and ambiguous objects ($F_{2,29} = 19.048$, $p < 0.0001$). Among the four sets, the only pair that is not significantly different is the improper-relation set and the ambiguous-object set.

The main result is that the difference in response time between a proper setting of the two objects (proper relation) and the same ambiguous object in isolation (single object) was significant. That is, the two objects were recognised better as a pair than one of them, the ambiguous object, alone.

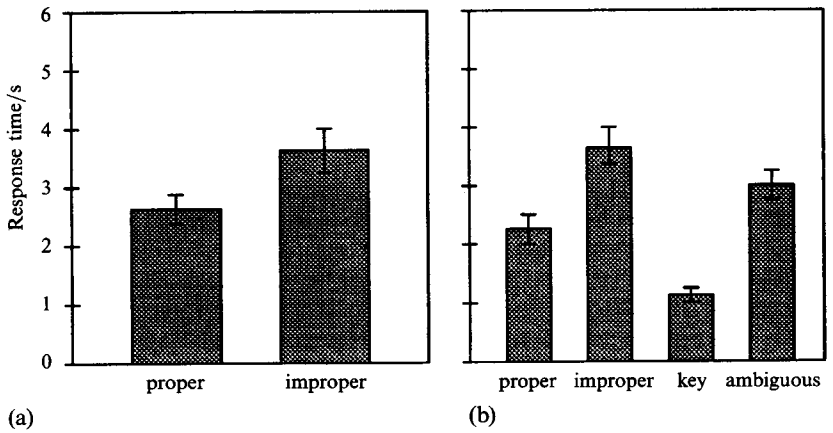

Figure 3. Response times for the different conditions. (a) A comparison between the recognition of objects with proper and improper spatial relations. (b) Performance in all four conditions (proper spatial relation, improper spatial relation, a single key object, and a single ambiguous object). Bars indicate standard error of the mean.

2.2.2 *Error rates.* The error rates for each stimulus, over all subjects, were calculated. For the comparison between error rates in the proper-relation and in the improper-relation sets a paired t-test was performed and revealed a significant difference ($t_{11} = 3.827$, $p < 0.003$). Among the fifteen pairs of proper-relation and fifteen pairs of improper-relation stimuli, twelve pairs involved the same key and ambiguous objects. In the proper-relation set subjects made an average of 29.7% errors, whereas they made an average of 49.0% in the improper-relation set (figure 4a). The mean difference was 19.3%.

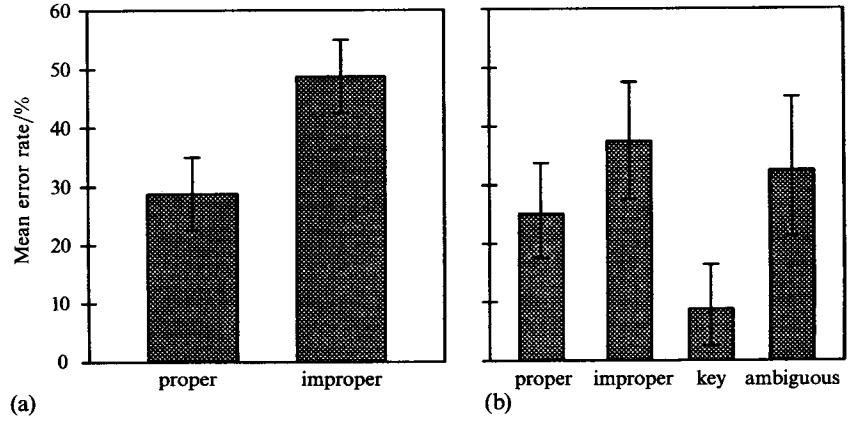

Figure 4. A comparison between error rates for recognition of objects for the different conditions (a) between proper and improper spatial relations; (b) between all four conditions (proper spatial relation, improper spatial relation, a single key object, and a single ambiguous object). Bars indicate standard error of the mean.

In section 2.2.1, we compared the cases where all the four versions (proper relation, improper relation, key object, and ambiguous object) were correctly identified. The corresponding error rates for these stimuli were an average of 24.8% errors in the proper-relation set, 36.9% in the improper-relation set, 6.9% in the key-object set, 35.2% in the ambiguous-object set (figure 4b).

Some of the incorrect responses to the isolated objects could be reasonable interpretations of these figures seen without context (for example, the face of the man in figure 1c in isolation could be interpreted as a brush). However, the correlation between the response times and error rates indicates that a time–accuracy trade-off strategy was not involved in the performances.

Two important results emerge from this experiment. First, the spatial organization between two objects has a significant effect on their recognition. Second, the recognition of two objects, key and ambiguous in proper spatial relation, is better than the recognition of only one of them, the ambiguous, in isolation. The implications of these results will be discussed in section 3.2.

3 Experiment 2: Ambiguous–ambiguous configurations

In order to examine further the role of a key object in the identification of a complex configuration, we added to the original stimuli set ten configurations that consisted of two ambiguous objects, rather than key and ambiguous objects (eg figure 5). The original spatial relations were maintained and therefore the new configurations can be considered as proper-relation configurations with an ambiguous object replacing the key object. The additional configurations consist of ambiguous objects that also appeared in other configurations in experiments 1 and 2, and their definition as ambiguous is thus supported by experimental evidence. Under conditions identical to the original experiment, seven subjects had to recognize both objects.

Figure 5. Examples of ambiguous–ambiguous configurations that were derived from figures 1a and 1b. All the ambiguous–ambiguous configurations were composed of two ambiguous objects with the original proper spatial relation.

3.1 *Results*

The results indicate a clear deterioration of recognition performance. While the worst results in the original experiment were 49.0% error rate and reaction time of 3.577 s, the results obtained in these configurations were 90.0% error rate and a mean reaction time of 6.65 s (figure 6).

On the average, the ambiguous–ambiguous condition was recognized with 90% errors, the proper-relation condition with 31%, and the improper-relation condition with 56%. The difference in performance between the ambiguous–ambiguous condition and both proper-spatial-relation and improper-spatial-relation conditions was highly significant. For differences in error-rate percentage the analyses yielded, $t_6 = 7.307$, $p < 0.00034$, for ambiguous–ambiguous vs proper relation; and $t_6 = 4.362$, $p < 0.0048$, for ambiguous–ambiguous vs improper relation (subjects did not produce enough correct response for analysis of the reaction time). The proper relation vs improper relation comparison was also significant in this experiment ($t_6 = 8.225$, $p < 0.00017$). The replacement of the key object by an ambiguous one had a critical influence on recognition performance.

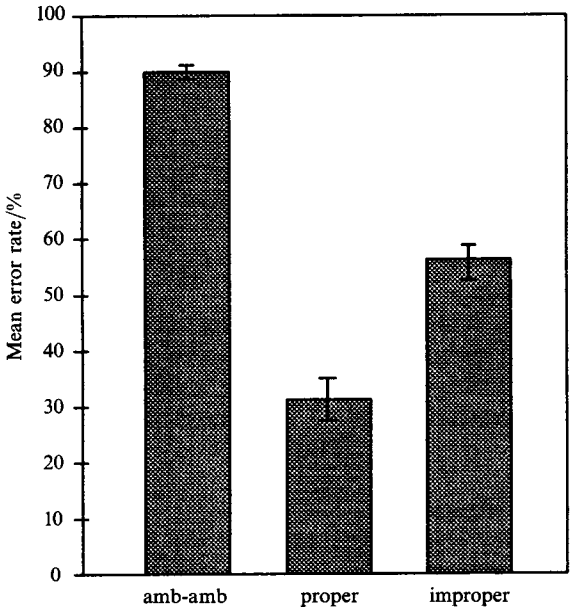

Figure 6. A comparison of error rates in recognition of the three different configurations of experiment 2. Two ambiguous objects in proper spatial relation (amb–amb, left), key and ambiguous objects in proper spatial relation (middle), and key and ambiguous objects in improper spatial relation (right). Bars indicate standard error of the mean.

3.2 *Discussion*

In the present study we were mainly interested in the interactions among objects in multiobject configurations, and in the relative contribution of objects of different identifiability to the global figure interpretation. In particular we were concerned with the following two questions. (1) What is the importance of spatial relations among objects in a scene, and how do they affect recognition performance? (2) How can the presence of a highly recognizable object in a scene influence the recognition of the complete figure?

The results show a substantial effect of spatial configuration on recognition performance. The analysis shows that subjects required considerably more time, and made more errors, when the spatial relations were improper compared with the proper-relation condition. Even when the subjects could identify the two constituting objects in isolation, the same trend, albeit somewhat weaker, was apparent. This result is consistent with results reported by Bar and Ullman (1993), Biederman (1972, 1981), Biederman et al (1982), Cave and Kosslyn (1993), Hock et al (1978), Mandler and Johnson (1976), and Palmer (1975a). Clearly identifiable glasses, for instance, can help the recognition of ambiguous face objects connected to them, but not (or help considerably less) when they are positioned in other locations in the scene. Consequently, recognition memory must contain information not only about the identity of objects that tend to co-occur in scenes, but also about their typical spatial relation. That is, the presence of a 'hat', say, in the image, can facilitate the recognition of a 'pipe', but the facilitation is not merely an association between the two categories, it depends also on relative location.

The second result emerges when we compare the statistics of the cases where all four versions of the same scene, proper relations, improper relations, key object, and ambiguous object, were correctly identified. From this comparison we conclude that the stimuli that were identified most readily and accurately were key objects, then came the proper relations stimuli, the ambiguous objects, and the improper relations stimuli.

(Recall that each scene, proper relations or improper relations, consisted of two objects, key and ambiguous, that also appeared in isolation at some other stage in the experiment.) In particular, it is noteworthy that the proper-relations condition that contained both an ambiguous and a key object was faster to recognize than only the ambiguous object in isolation. Although multiobject configurations have been previously explored, the effect of the identifiability of one object on the recognition process of another object was not tested directly [for example, the results reported in Palmer (1975a) demonstrate the effect of a scene on the recognition of subsequently presented objects]. The results reported here suggest that the ambiguous identity of an object can be resolved by the presence of a related and clearly identifiable object. The ambiguous–ambiguous experiment revealed how crucial is the absence of the key object. It seems reasonable to assume that when there is no 'triggering object' for recognizing such scenes, recognition is degraded. Along the same line, we may expect that the improper-relations set is more difficult to recognize than the ambiguous-objects set because it requires the identification of two separate items, and at the same time it provides no help and may sometimes provide misleading clues.

3.3 *Implications for the organization of recognition memory*

The idea that mental representations are influenced by associations has a long tradition, dating back to the British empiricists, including Locke, Berkeley, and Hume. The empiricists suggested that ideas and impressions are associated by their tendency to co-occur.

A simple form of using associations between objects for the purpose of recognition would be to link together in recognition memory objects that tend to co-occur, as in semantic networks (Quillian 1968), associative memories (Kohonen 1984), and in typical-scenes schemata (Palmer, 1975b). Our results suggest a more complex organization of recognition memory that goes beyond linking related objects, and stores in addition information about their typical spatial relations.

A possible suggestion is that objects are organized in recognition memory in structures that depict typical scenes. Such a structure, which may be called a 'context frame' contains a number of objects such as a face, a hat, glasses, in a typical configuration. A given object may appear in more than a single context frame. During recognition, an object can select a context frame (or a set of frames), and a frame can select an object (or a set of objects). When an object is recognized, it invokes context frames in which it appears. The frames then set expectations not only about other possible objects in the scene, but also about their expected location, scale, and orientation. In recognition, the goal is to determine the identity of viewed objects, despite possible variations in position, scale, orientation, etc (Grimson 1990; Lowe 1986; Ullman 1989). Information derived from the context frame regarding the expected identity of other objects, as well as their position, orientation, scale, etc, could therefore facilitate significantly the recognition of related objects.

As a consequence of this study, we can conclude that a key object may serve as a trigger that enables the recognition of other objects in the scene. It appears that a context frame is typically invoked by the recognition of at least one object contained in the frame. We had no evidence to suggest that it can also be invoked by the cooperative activity of multiple ambiguous objects.

The interplay between the recognition of individual objects and the invocation of context frames can account in part for the efficiency of the human visual system in interpreting natural scenes, including complex scenes containing multiple objects, some of which may be partially occluded or poorly visible. The incorporation of similar mechanisms in artificial recognition systems, rather than attempting to recognize each object in the scene individually, could also provide a crucial step in improving their capacity to deal with natural scenes.

Acknowledgements. We are indebted to S Edelman, M Lando, R Basri, E Schechtman, and the Vision Group at the Weizmann Institute of Science for significant contribution to this work. We thank R T Green and M C Courtis for the use of figure 1. Parts of this work were presented as a poster at the Association for Research in Vision and Ophthalmology meeting 1995. This work was supported in part by the Israel Science Foundation administered by the Israel Academy of Sciences and Humanities.

References

Bar M, Ullman S, 1993 "Spatial context in recognition", technical report CS93-22, The Weizmann Institute of Science, Rehovot, Israel

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

Biederman I, 1981 "On the semantic of a glance at a scene", in *Perceptual Organization* Eds M Kubovy, J Pomerantz (Hillsdale, NJ: Lawrence Erlbaum Associates) pp 213–253

Biederman I, Mezzanote R J, Rabinowitz J C, 1982 "Scene perception: Detecting and judging objects undergoing relational violations" *Cognitive Psychology* **14** 143–177

Cave C B, Kosslyn S M, 1993 "The role of parts and spatial relations in object identification" *Perception* **22** 229–248

Green R T, Courtis M C, 1966 "Information theory and figure perception: The metaphor that failed" *Acta Psychologica* **25** 12–36

Grimson W E L, 1990 *Model-Based Vision* (Cambridge, MA: MIT Press)

Hock H S, Romanski L, Galie A, Williams C, 1978 "Real-world schemata and scene recognition in adults and children" *Memory and Cognition* **6** 423–431

Kohonen T, 1984 *Self-organization and Associative Memory* (Berlin: Springer)

Lowe D G, 1986 *Perceptual Organization and Visual Recognition* (Boston, MA: Kluwer-Nijhoff)

Mandler J M, Johnson N S, 1976 "Some of the thousand words a picture is worth" *Journal of Experimental Psychology: Human Learning and Memory* **2** 529–540

Palmer S E, 1975a "The effects of contextual scenes on the identification of objects" *Memory and Cognition* **3** 519–526

Palmer S E, 1975b "Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction", in *Explorations in Cognition* Eds D A Norman, D E Rumelhart (Hillsdale, NJ: Lawrence Erlbaum Associates) pp 297–307

Quillian M R, 1968 "Semantic memory", in *Semantic Information Processing* Ed. M Minsky (Cambridge, MA: MIT Press) pp 227–270

Ullman S, 1989 "Aligning pictorial descriptions: an approach to object recognition" *Cognition* **32** 193–254