

Spatial Context in Recognition

Moshe Bar Shimon Ullman

Department of Applied Mathematics and Computer Science,
The Weizmann Institute of Science,
Rehovot 76100,
Israel.

December 18, 1993

Abstract

In recognizing objects and scenes, partial recognition of objects or their parts could be used to guide the recognition of other objects. Here, we investigated psychophysically the role of local features in the recognition of complete figures, and the influence of contextual information on the identification of ambiguous features. We used configurations of features that were placed in either proper or improper spatial relations, and measured response times and error rates in a recognition task. Two main results were obtained. First, proper spatial relations among the features of a scene decrease response times and error rates in the recognition of individual features. Second, the presence of features that have a unique interpretation in the scene disambiguates the identity of ambiguous features faster and with less errors compared with the same features that appear in isolation, or in improper spatial relations. The implications of these findings to the organization of recognition memory are discussed, and a framework for a model for using spatial context, which uses the psychophysical findings, is proposed.

1 Introduction

1.1 Spatial Context

Natural scenes usually contain multiple objects, and different scenes (such as a street, an office, etc.) are associated with different groups of characteristic objects.

Given a certain context, we can immediately name possible members of the scene in question. Furthermore, one is able to expect not only the scene constituents, but also their probable location in the scene and the possible spatial relations among them (e.g., a hat is more probable to appear above a person's head than in other locations in the scene). One explanation for the efficiency of the recognition process in certain circumstances (such as those depicted in fig. 1) may be by means of *Spatial Context*. Partial recognition of an object's part, or the identification of a single object within a configuration of several objects, can facilitate the recognition of other parts and other objects in the scene. For example, an ambiguous blob may be interpreted as a head when we recognize a hat above it.

It is still unclear how spatial context aids visual recognition in human perception, and current recognition systems usually do not make use of spatial context information. In the present study we investigate certain aspects of spatial context in human vision, and propose a general framework for using spatial context in computer vision.

1.2 Previous studies

The recognition of scenes with different types of inter-object arrangements was examined in a number of experiments. Hock *et al.* (1978) investigated recognition memory and found that organized scenes with possible inter-object relations are recognized more accurately than unorganized scenes with impossible inter-object relations. Subjects were presented with pairs of scenes, one of which had been seen previously, and had to identify the scene that was presented in the initial phase. Coherent Scenes were easier to identify than incoherent arrangements. The study did not test directly, however, the recognition of individual objects.

Palmer (1975a) examined the influence of prior presentation of visual scenes on the identification of briefly presented drawings of real-world objects. Presentation of a scene provided a context (or no context) for a subsequent target object, which could be either consistent or inconsistent with the specific context. The probability of identifying the object correctly was higher in the appropriate context conditions. Although the experiment did not examine directly the effect of spatial relations on recognition performances, it emphasized the role of context in identification tasks.

The perceptual processing of scenes with different inter-object organizations was comprehensively investigated by Biederman and his colleagues. Biederman (1972) presented subjects with jumbled real-world scenes and discovered that the accuracy of identifying a single cued object was lower than when the presented scene was coherent. The jumbling was obtained by cutting photographs of real-world scenes into equal sections and mixing them, so that they formed new jumbled scenes.

The subjects had to indicate, by pointing to one of four object pictures that were presented either before or after the scene, which object had been cued. When the objects were presented before the scene, the task can be considered a visual search task. The main difference between search and recognition tasks is that when one is searching for an object, a certain object-model is known, a priori, only its location must be determined. In recognition, the object model, and sometimes also its location, 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 the search task, where we have a clear idea as to what we are looking for. Even when the objects were presented after the scene, the subjects were not required to name the objects, and they could conceivably be correct without knowing the identity of the object.

Biederman assumed that the jumbling was "a manipulation of the meaningfulness of the object's setting independent of the complexity of the scene". However, in addition to the spatial relations and the context, the original real-world scene is rich with information such as shapes, shading, textures etc. Hence, in addition to the jumbling of the spatial relations, abnormal discontinuities of the scene properties arise and it is difficult to assess their effect on the overall performance. Biederman defined scene semantics by defining five classes of relations that are needed to characterize much of the organizational difference between a coherent scene and a scene containing unrelated objects (Biederman 1981): *Support* - objects are supposed to rest on appropriate surfaces, *interposition* - opaque objects occlude other objects, *probability* - objects have a certain probability of appearing in a certain scene, *position* - once appearing in a scene, objects have a certain probability of appearing in a certain position in that scene, and *size* - objects have a particular size relative to other objects in the scene. In a number of experiments subjects were briefly presented with scenes in which an object in a cued location either conformed to the normal relation with its background or violated one of the relations (Biederman 1981, Biederman *et al.* 1982). (For example, a hydrant on a mail box violates the position relation.) The detection of objects in an abnormal setting was less accurate and slower than when the same objects maintained normal relations with their environment. As the number of violations increased, target detectability generally decreased. The subjects in these experiments were first provided with the target object names, and then they had to detect (locate) them in the scene. As explained above, these are also close to search rather than recognition tasks, but they nevertheless provide us useful information concerning scene perception.

In the experiments, subjects had to determine whether a target object was present in a display of objects that belong to the same possible scene, but were arranged in a "nonscene" manner. The target objects could be either consistent or inconsistent with the setting (the other objects in the display). In contrast to expectation, no apparent benefit was observed to for targets that were likely to appear

with the setting (i.e., consistent with the setting). The lack of proper spatial relations was the main difference between these experiments and other experiments that examined the role of contextual consistency (and yielded an advantage in performance to the proper context conditions). It can be inferred that, in addition to the role of context, spatial relations add a significant contribution to the perception of multi-object scenes. Mandler and Johnson (1976) tested the recognition of objects in scenes, using distractors that varied the organization of the objects collections (spatial relations, the number of objects contained in the picture, figurative detail of the objects, etc.). Their results also indicate that spatial location is a major cue in scene recognition.

Context can also affect lower levels of visual processing. Weisstein and Harris (1974) found that context can influence orientation discrimination. They presented subjects with four line segments differing in orientation and location relative to a fixation point. On each trial, one of these target lines, together with one of several context patterns, was flashed briefly, producing a compound pattern. The main finding was that when a target line was a part of a configuration that looks coherent and three dimensional, it was identified more accurately than in other contexts. In a related experiment, Wong and Weisstein (1982) used figure-ground ambiguous pictures and found that a barely visible, briefly flashed line segment, was discriminated more accurately when it was a part of an object-like pattern, than when it was flashed alone, or when it was a part of a random collection of lines. Cave and Kosslyn (1993) investigated the role of object's parts and their spatial relations in single object identification. They used drawings of individual objects as stimuli, and found that proper spatial relations among the object's components were important for fast and correct identification. They also found that the way objects were divided into parts has a little affect on recognition, the perceptual abilities is evident given that the original spatial relations were preserved. An application that demonstrates the aid of spatial relations information to recognition is described in a computational study by Srihari and Rapaport (1989). They developed a system that used textual clues about spatial relations, as they appear in a captions of newspaper photographs, to infer the identity and spatial relations of visual objects in the photographs.

The findings reviewed above highlight the importance of context and spatial relations for visual perception and recognition. However, most of the studies that examined scene perception were mainly concerned with visual search and detection, rather than recognition per se. Furthermore, previous studies did not investigate explicitly the influence of the presence of a highly recognizable object on the disambiguation of an ambiguous object. Our goal in the present study was therefore twofold. First, we wanted to examine the effect of the spatial relations between two objects, that originally belonged to the same figure, 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. As we shall see in the discussion, this question has important ramifications to the organization of recognition memory. Second, we wanted to examine how an object which has a unique interpretation can disambiguate the identity of a more ambiguous object.

2 Spatial Context: psychophysical experiments

The main goal was to test the effect of recognizing one part in a configuration on the recognition of related parts, and the dependence of such an effect on spatial arrangement. We also tested the influence of a “key feature”, that has a clear unambiguous interpretation, on the recognition of the complete figure.

2.1 *Method*

2.1.1 *Subjects*

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

2.1.2 *Materials*

63 stimuli were derived from the eight figures depicted in fig. 1. Three types of stimuli were created: isolated features (F), two features that maintained their proper spatial relations (P), and two features with improper spatial relations (IP). The two features that composed a scene, either (P) or (IP), were chosen in such a way that one of them was easier to identify (“Key-feature”), while the second was more difficult to identify (“Ambiguous-feature”). (Our initial selection of key and ambiguous features was later supported by means of response times and error rates). The spatial relations were scrambled by placing the features in random locations, keeping approximately the original physical distances (fig. 2).

Since the order in which the stimuli appear may affect the results (e.g., the performances of a subject on recognizing (IP) stimuli may differ when they appear before or after the (P) stimuli of the same figure), we ordered the stimuli so that the (P), (IP), and (F) stimuli of a certain figure were presented in a different order in each test set. Hence, 63 stimuli, which consisted of 33 (F)’s, 15 (P)’s, and 15 (IP)’s, were ordered in six different ways (all six combinations of ordering (F), (P), and (IP) versions), forming six different test sets.

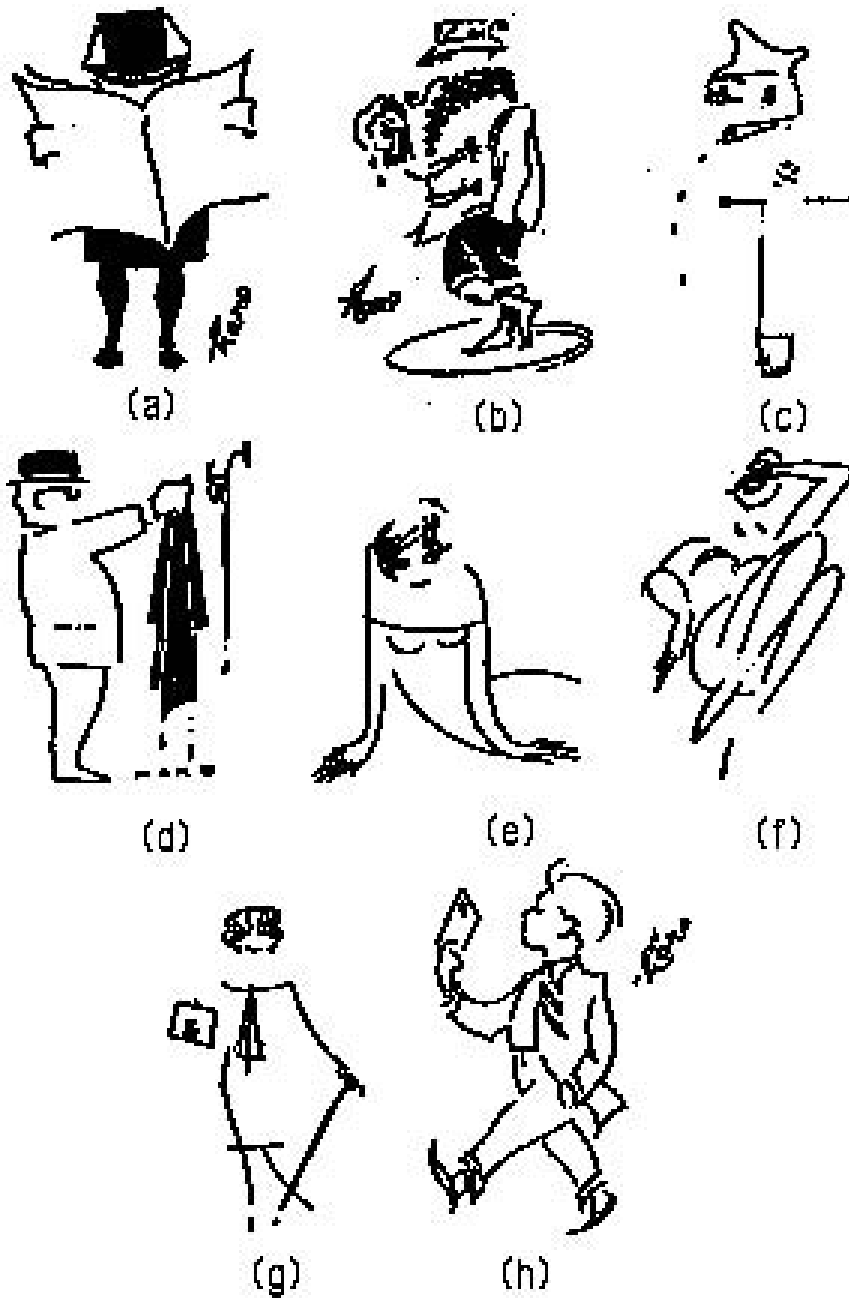


Figure 1: Examples of the original figures from which we produced our stimuli. [Reproduced with permission from Green, R. T., & Courtis, M. C. Information theory and figure perception: The metaphor that failed. *Acta Psychologica*, 1966, 25, 12-36]. One can notice how figures that are consisted of identifiable, ambiguous and missing features can be meaningful.



(a)



(b)



(c)



(d)

Figure 2: Examples of the different stimuli. (a) Key-feature, (b) Ambiguous-feature, (c) Proper spatial relations, and (d) Improper spatial relations. Subjects had to identify all objects in each stimulus.

2.1.3 *Apparatus*

The drawings were scanned by a MICROTEK scanner and manipulated by a Macintosh *IIfx* computer. A Silicon Graphics Personal Iris 4D/35 computer, with a 1280 x 1024 resolution screen, controlled the stimulus presentation, and recorded the response times (RT) and the error rates.

2.1.4 *Procedure*

The subjects were tested individually. Prior to testing, they were given instructions requiring them to name each object in the stimuli aloud as quickly and as accurately as possible. The subjects were asked whether they had any questions and all questions about the procedure were answered at this time.

Each subject viewed one of the six sets of 63 stimuli. The subjects sat approximately 40 cm in front of the computer screen. The stimuli subtended an area approximately 9.5 cm x 6.5 cm (6.77 deg x 4.65 deg of visual angle) 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 following the responses were recorded by the experimenter. Stimuli presentations were separated by a 3 s interval.

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 (we had only one subject whose responses had to be omitted); an outlier was defined as response time that was greater than 3 times the mean of that condition without the outlier. We also calculated the number of errors made in each condition. An erroneous response was defined as an interpretation that was different from the unique interpretation that is assigned to the original figure, or when the subject was unable to produce any interpretation. (Although other definitions of a correct and error responses can be used, this definition meets our goals). The names of higher-level categories were considered correct (e. g., ‘man’ when a soldier was shown). The mean response time for correct responses was 2.36 s, and the mean error rate was 26 %. All results not reported were not significant ($p > 0.05$).

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 that correctly identified the (P) and its corresponding (IP) version (48 cases). The comparison between performances in both these versions is depicted in fig 3(a). A paired (differences) t-test was performed on these data. The difference in mean response times was significant. Scrambling the spatial relations greatly increased naming times: features that were presented in (IP) spatial relations were correctly identified in an average of 3.535 s, whereas features that were presented in (P) spatial relations were correctly identified in an average of 2.565 s. The mean difference was 0.97 s ($t_{47} = -2.710, p < 0.009$).

We then considered the responses of subjects that not only respond correctly to the (P) and (IP) versions, but also correctly identified the two stimuli of isolated features which constituted the two-objects scene (30 cases). The comparison between performances in these four versions is depicted in fig 3(b). The set of (P) stimuli was correctly identified in an average of 2.183 s, the set of (IP) in an average of 3.577 s, the set of “key-features” in an average of 1.357 s, and the set of “ambiguous features” in an average of 3.051 s. An analysis of variance was performed on these data and it revealed a significant difference between the response times of correctly identifying (P), “key-features”, and “ambiguous-features” ($F_{2,29} = 19.048, p < 0.0001$). Among the four sets, the only pair that is not significantly different is the (IP) set and the “ambiguous feature” set. The main result is that the difference in response time between a proper setting of ambiguous and key features, (P), and the same ambiguous feature in isolation, (F), was significant. That is, the two features are recognized better together than the ambiguous feature alone.

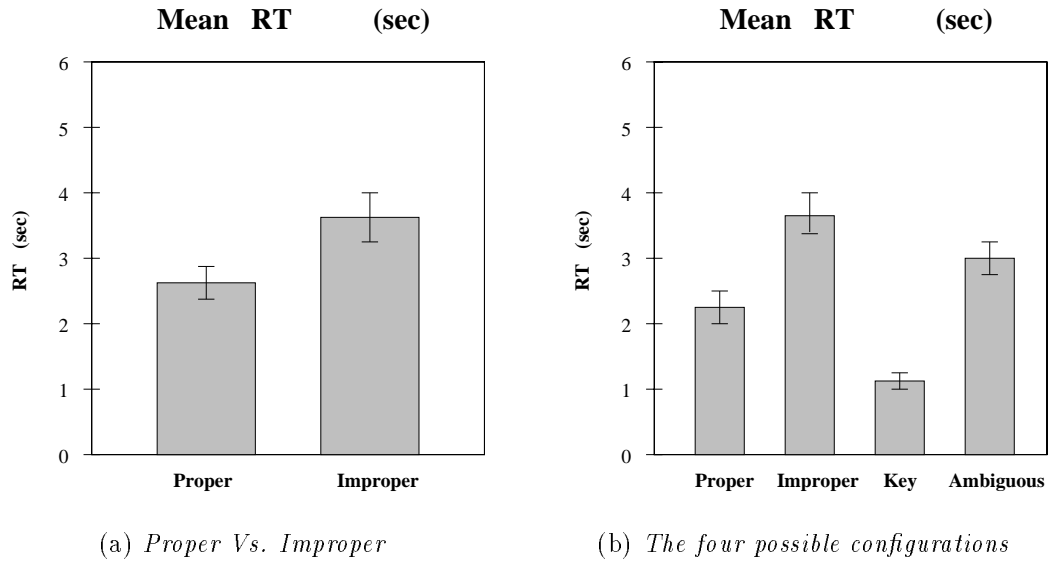


Figure 3: Comparison between RT for different spatial relations

2.2.2 Error rates

The error rates for each stimulus, over all subjects, were calculated. For the comparison between error rates in the (P) and in the (IP) sets a paired t-test was performed and revealed a significant difference ($t_{11} = 3.827, p < 0.003$). In the (P) set subjects made an average of 29.7 % errors, whereas they made an average of 49.0 % in the (IP) set (fig. 4(a)). The mean difference was 19.3 %.

The mean error rates, for the analysis of the data where all four versions ((P), (IP), “key-feature”, and “ambiguous-feature”) were correctly identified (fig. 4(b)), are: average of 24.8 % errors in the (P) set, average of 63.1 % errors in the (IP) set, 6.9 % average of errors in the “key-feature” set, and 35.2 % in the “ambiguous-feature” set.

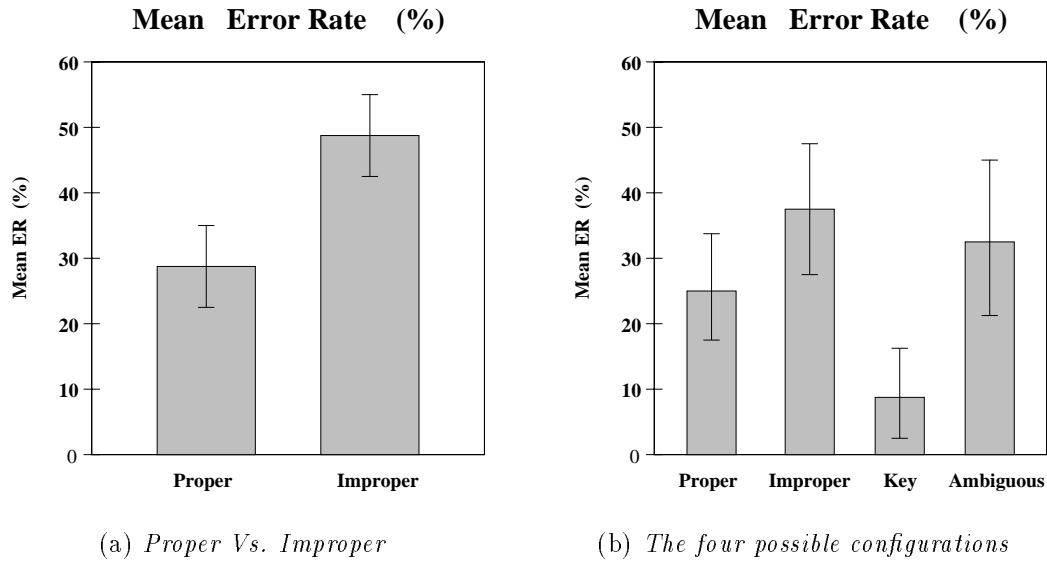


Figure 4: Comparison between error rates for different spatial relations

2.3 Discussion

The results show a substantial effect of spatial configuration on recognition performances. The analysis shows that subjects required considerably more time, and made more errors, when the spatial relations were improper compared with the proper condition. Even when the subjects could identify the two constituting features in isolation, the same trend, albeit somewhat weaker, was apparent. It is important to note that context effects depended not only on the identity of the constituent parts, but also on their spatial arrangement. That is, the presence of a ‘hat’, say, in the image, can facilitate the recognition of ‘glasses’, 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, (P), (IP), “key-feature”, and “ambiguous-feature”, were correctly identified. From this comparison we conclude that the stimuli that were identified most readily and accurately were those of “key-features”, then came the (P) stimuli, the “ambiguous-features”, and the (IP) stimuli. (Recall that each scene, (P) or (IP), consisted of two features, key and ambiguous, that also appeared in isolation at some other stage in the trial.) In particular, it is noteworthy that the (P) condition that contained both an ambiguous and a key feature, was faster

to recognize than the same ambiguous feature in isolation. A natural explanation for this result is that the presence of the “key-feature” in the (P) scene helped the identification of the “ambiguous-feature”, that was identified more slowly when it appeared without context. Along the same line, we may expect that the (IP) set is more difficult to recognize than the “ambiguous features” 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.

The correlation between the response times and error rates indicates that a time-accuracy trade-off strategy was not involved in the performances.

3 General discussion

3.1 Relation to previous results

The experiment was designed to study the recognition of multi-object configurations. In particular, we asked how the identification of one object contributes to the recognition of the another object. The findings of this research are in general agreement with the outcome of related studies, but some differences are noteworthy. Our first result was that objects in organized settings are recognized faster, and more accurately, than objects in settings where the spatial relations are unnatural. This result is consistent with the results reported by Biederman (1972, 1981), Biederman *et al.* (1982), Cave and Kosslyn (1993), Hock *et al.* (1978), Mandler and Johnson (1976), and Palmer (1975a). One difference between the current and previous studies is that most of the previous experiments, as described in the introduction, explored performances in search and detection task whereas the current experiment examined performances in a recognition task, where the subjects did not have prior knowledge about the target objects. The recognition task is more probably difficult, (the mean RT in this experiment is usually longer than the mean RT in related search tasks), and the cognitive mechanisms involved may be different than those subserving the search tasks. 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, Biederman 1981, Mandler and Johnson 1976, Palmer 1975a), our configurations consisted of two separate objects. This allows 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.

The second result was that the recognition of two objects, the key and ambiguous features together, was faster and more accurate than the recognition of the ambiguous feature alone. Although inter-object configurations have been previously explored, the effect of the identifiability of one object on the recognition process of

another object, was not tested directly (e.g., the results reported in Palmer (1975a) demonstrate the effect of a scene on the recognition of subsequently presented objects). The result reported here suggest that the ambiguous identity of an object can be resolved by the presence of a clearly identifiable object.

3.2 Implications for the organization of recognition memory

The most significant result of the present study is that the interaction between objects during recognition depends not only on the identity of the objects but also on their spatial arrangement. Clearly identifiable glasses, for instance, can help the recognition of ambiguous face features connected to it, but not (or considerably less so) 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 relations.

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), in associative memory (Kohonen 1984), and in typical scenes schema (Palmer 1975b). Biederman (1981) described the possible use of a scene schema that could be directly accessed by a single object identification, and the information contained in this schema can then set expectations regarding the identity of other objects in the scene. Our result suggest a more complex organization 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. We call such structures “context frames”. A context frame contains a number of objects such as a face, a hat, glasses, etc., 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 spatial context frames in which it appears. The frames then set expectations not only about other possible objects, 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 1985, Ullman 1989). Information derived from the context frame regarding the expected identity of other objects, as well as their position, ori-

entation, scale, etc., could therefore facilitate significantly the recognition of related objects.

Current artificial recognition systems typically use a set of 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 1985, Ullman 1989). Using context information in the manner suggested here could help object recognition in multi-object configurations to proceed more efficiently, and cope with degraded, ambiguous, or missing information in the image. We are currently exploring this possibility by computer simulations.

Acknowledgments

We would like to thank S. Edelman for fruitful discussions and assistance with the experiments and E. Schechtman for help in the statistical analysis.

References

- Biederman I. (1972). Perceiving real-world scenes. *Science* Vol 177, 77-80.
- Biederman I. (1981). On the semantics of a glance at a scene. In M. Kubovy and J.R. Pomerantz (Eds.) *Perceptual organization*. Hillsdale, N. J.: Erlbaum.
- Biederman I., Blicke T.W., Teitelbaum R.C., and Klatsky G.J. (1988). Object search in Nonscene Displays. *Journal of Experimental Psychology: Learning, Memory, and Cognition* Vol. 14, No. 3, 456-467.
- Biederman I., Mezzanotte R.J., and 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* Vol. 22, 229-248.
- Grimson W.E.L. (1990). Object recognition by computer: The role of geometric constraints. Cambridge, Mass. MIT Press.
- Hock H.S., Romanski L., Galie A., and Williams C.S. (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. Springer-Verlag.
- Lowe D.G. (1985). Perceptual organization and visual recognition. Kluwer Academic, Boston.
- 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.
- Quillian M.R. (1968). In Semantic information processing. Minsky M. (Ed.). Cambridge, Mass. MIT Press.
- Srihari R.K., and Rapaport W.J. (1989). Combining linguistic and pictorial information: using captions to interpret newspaper photographs. In

Current Trends in SNePS - Semantic Network Processing System.

Palmer S., (1975a). The effects of contextual scenes on the identification of objects. *Memory and Cognition* 3:519-526.

Palmer S., (1975b). Visual perception and world knowledge: notes on a model of sensory-cognitive interaction. In *Explorations in Cognition*, Norman D.A. and Rumelhart D.E. (Eds.). Freeman, San Francisco.

Ullman S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32:193-254.

Weisstein N., and Harris S. (1974). Visual detection of line segments: an object-superiority effect. *Science* 186:752-755.

Wong E., Weisstein N. (1982). A New perceptual context superiority effect: Line segments are more visible against a figure than against a ground. *Science* 218:587-589.