What do feature detectors detect?
Features that encode context and the binding problem

Robert Ward

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
ward@psy.cmu.edu

Abstract

The representation of visual features is investigated by examining the types of information that are encoded at the feature level which are used for feature binding. Features are often assumed to be bound together by virtue of their common location, but the current study shows that shared context, as well as location, acts to constrain the feature binding process and the formation of illusory conjunctions. Two different sorts of context manipulations are reported. In one manipulation, the context of each item in the display is established by flanking bars, and binding errors are examined as a function of this shared context. Also examined is a more global context manipulation in which the items presented form either a word or nonword. Both sorts of contexts affect feature binding, although in different ways. Finally, some of the computational difficulties in implementing a feature representation that encodes context are considered.

Does a feature detector detect something more than just the presence of a feature? In fact, the answer is probably yes. For example, consider a retinotopic feature map, in which many feature detectors with limited receptive fields are distributed to span a large region of visual space. Since each detector is tied to a particular location within the visual field, an active detector specifies both the presence of a feature and the location of that feature. In such a feature map, it would be correct to say that features are encoded with location information.

The possibility that feature detectors may encode more than the mere presence of a feature is important when considering the problem of *feature binding* -- that is, when determining which features belong to the same object and which features belong to different objects. Figure 1 shows a set of feature detectors that encode only the presence of features. The activation of these detectors by themselves cannot be unambiguously interpreted. If instead we have a set of retinotopic feature maps, and we know the correspondence between locations on the two maps, then the activity of the feature detectors specifies not only the location of features, but also the location of feature conjunctions. Theorists in psychology have differed on how easily the correspondence between two maps might be accessed, or how available this correspondence might be. For example, Treisman and Gelade (1980) assumed that focal attention was required to determine the features present at any single location; more recently, Wolfe, Cave, and Franzel (1989) allowed that in effect, feature maps could be superimposed upon each other, so that the combined activity of relevant feature maps could be determined for all locations simultaneously. But what is important to note is that these different theories share the idea that features are bound together by virtue of their common location.

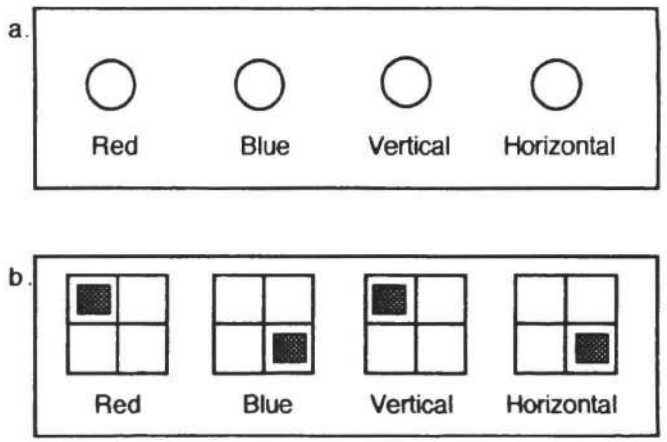

Figure 1. a) This feature representation cannot distinguish between the simultaneous presentation of a red vertical with a blue horizontal line and a red horizontal with a blue vertical line. b) This representation specifies the location of all features as well as the location of the red vertical and blue horizontal lines.

But should location be special? Or might features encode other types of information? Might there be a way of identifying which features belong together besides shared location? Duncan (1989) suggests that the visual system is most likely not limited to location cues for integrating features together. This report will examine the possibility that features encode aspects of their *surrounding context*. In this way, features that share a similar context would be bound together in the same way as features sharing similar locations. For example, consider the letter A in the word CAT. One constraint on the way the features of the A

would be conjoined is their shared context: all the features of the A appear in a context consisting a C on one side and a T on the other. This information could be useful in deciding that all the features of the A belong together.

It is already well-established that context plays a central role in our perceptions. The famous example illustrated in Figure 2 shows how the same figure, in this case the ambiguous A/H object, can be interpreted in different ways depending upon the object's surround. However, it is a very large jump from the claim that context can affect our interpretation of objects, to the claim that aspects of context are encoded at the feature level. Furthermore, there is a considerable difference between encoding location information at the feature level and contextual information at the feature level.

Figure 2. An example of the effect of context on the interpretation of an ambiguous object.

It seems intuitively reasonable, or at least consistent with the known anatomy of the visual cortex, to imagine a feature representation consisting of many varieties of retinotopic maps. Such maps can be quite general purpose, in that they can specify all the possible locations in which an object might appear. It seems much more difficult to specify all the different *contexts* in which an object might appear; in fact, it can be hard to visualize the organization of a feature "map" encoding contextual information. Figure 3 shows one possibility. The map specifies the presence of different features within several contexts. Contexts which are similar to each other are close together in this contextual encoding space. Such a map has some interesting properties. The contextual map does not specify *where* in the image a feature may occur, but instead specifies *what* sort of surround the feature occupies. Although this map does not directly encode location, it is still useful for feature binding, since features that share the same contexts will tend to belong together. Also, such a map can explain some sorts of perceptual grouping effects. In the representation used in Figure 3, a set of colinear points will share a more similar representation than a set of haphazardly arranged points. This similarity in encoding could serve as one basis for grouping the points together, so that the perceptual grouping would be explained at the level of feature representation.

The experiment reported here is based upon a principle of encoding similarity. It is assumed that the processes responsible for binding together features use information encoded by the feature detectors to determine likely feature conjunctions. Features that are encoded in similar ways, perhaps due to their similar locations or similar contexts, are therefore likely to be bound together. Likewise, illusory conjunctions, or the false recombinations of features (Treisman & Schmidt, 1982), should be more likely to occur among features given similar encodings. Cohen & Ivry (1989) found higher rates of illusory conjunctions among features that were displayed close together, suggesting that features encode location information. In the same way, if contextual information is encoded at the feature level, then illusory conjunctions should occur more frequently between features sharing similar contexts.

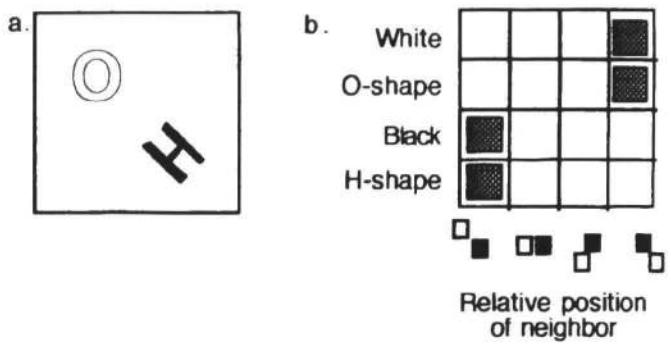

Figure 3. A sample feature map based on surrounding context. a) Sample stimulus. b) Encoding of stimulus using context based feature map. Each detector encodes information about the context of its associated feature, in this case the position of neighboring elements.

Two sorts of contexts are examined in the current experiment. The first sort of context is very simple, consisting of a pair of bars that flank display items either vertically or horizontally. This context manipulation, referred to here as "shared context", is local to the features present in the display; for example, items in the display can be presented in the same or in different contexts, and the effect of shared local context on feature binding can be observed. Based on the principle of encoding similarity, features presented in the same contexts should be more likely to be incorrectly conjoined. This simple context manipulation seems like a good starting point for finding contextually bound representations. The position of neighboring elements has been suggested before as a nonspatial basis for feature binding by Strong and Whitehead (1989). In the Strong and Whitehead model, the relative position of elements outside the focus of attention are used to provide a unique "tag" for the set of features within the focus of attention. The second type of context is a more global property of the stimulus, namely whether the features presented comprise a word or a nonword. An effect of these global properties would imply that the feature representation is determined only after a higher level of processing for which the global properties can be defined. At the same time, the experiment further investigates the encoding of location information by systematically varying the distance between features in these contexts.

Method

Stimuli. The experimental materials used in this experiment were very similar to those developed by Prinzmetal and Millis-Wright (1984, Experiment 2). Each stimulus display contained a string of three colored letters that appeared unpredictably to the left or right of fixation. There were six possible strings formed by the three letters, all containing the letter P: the words PIE, SPY, and MAP, and the nonwords PLF, BPT, and NVP. Letters were randomly colored in red, blue, green, or yellow, with the constraint that the same color was never used more than once in a letter string.

Each letter appeared in either a vertical or horizontal context. The vertical context consisted of white bars appearing above and below the letter; the horizontal context consisted of bars to the left and right of the letter. The bars forming the context always appeared to "bracket" their associated letter - that is, bars in the vertical context were oriented horizontally and bars in the horizontal context were oriented vertically. Stimuli were constrained so that each display contained letters in both the horizontal and vertical contexts, for a total of six possible arrangements of contexts among the letters. Figure 4 illustrates a sample display.

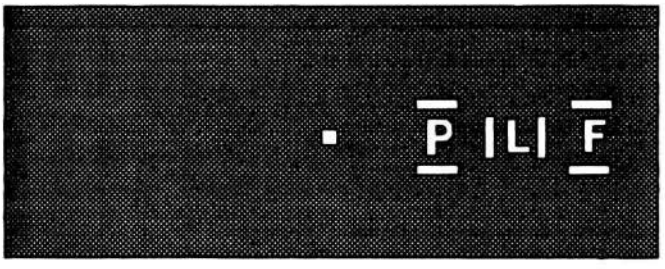

Figure 4. Sample stimulus. Each letter subtends 0.8 x 0.8 degrees with 2.4 degrees between letters. In this example, the P and F appear in the vertical context, and the L in the horizontal context.

Procedure. The task for subjects was to decide if the target letter P was the same color as a predesignated signal color. Each trial proceeded as follows. A colored square appeared in the center of the display for 1000 msec. The square served both to mark the fixation point and to designate the signal color for the trial. The stimulus display was then presented for 150 msec. Subjects pressed a key to indicate whether the color of the target letter P matched or did not match the signal color. Subjects were instructed to answer as quickly as possible without sacrificing accuracy. However, no feedback was provided during the experiment.

There were a total of 216 trials in the experiment. Pairing each of the six letter strings with each of the six possible context arrangements yielded 36 stimulus items, which appeared both to the left and right of fixation for a total of 72 stimulus displays. Each display was presented three times: once as a positive trial, in which the color of the target letter P matched the signal color; once as a feature trial, in which the signal color was not present in the display; and once as a conjunctive trial, in which the signal color appeared in one of the letters other than the target letter P. The conjunctive trials were the crucial trials of the experiment, since they allowed the possibility of feature conjunction errors. In all, three variables were manipulated in the conjunctive trials of the experiment: (1) whether the display consisted of a word or nonword; (2) the distance in characters between the letter P and the letter appearing in the signal color; and (3) whether the P and the signal color shared the same horizontal or vertical context. Subject to these constraints, the color of stimulus materials was determined randomly.

Before the experiment began, subjects received 32 practice trials in which feedback was provided.

Subjects. Fourteen Carnegie Mellon undergraduates participated for class credit. The data from one subject were discarded due to very high error rates, averaging 48% correct on the conjunctive trials, which are the crucial trials of the experiment, and 57% over all trials.

Results

The average error rates for positive trials, feature trials, and conjunctive trials were 34%, 6%, and 23% respectively. There were significantly more false alarms in the conjunctive conditions than in the feature conditions, $t(12) = 44.5$, $p < 0.001$. False alarms on the conjunctive trails are assumed to indicate feature binding errors; however, clearly there may be other causes for errors on the conjunctive trials. Therefore, it is not the absolute levels of errors that are interesting, but the effects of word stimuli, distance, and shared context on performance in the conjunctive trials. These effects are illustrated in Figure 5. A rich variety of both spatial and nonspatial influences on feature representation and binding can be examined in the current experiment: the effects of a simple shared context, the effects of distance between the conjoined features, and the effects of a "high level", more global context, created by word versus nonword stimuli.

The effects of shared context will be examined first. Averaging over the word and distance factors, subjects were significantly more likely to make conjunctive errors on the shared context trials than on the different context trials. False alarm rates averaged 27% for shared context trials and 20% for different context trials, $F(1,12) = 5.225$, $p = 0.041$. The probability of subjects conjoining features is therefore affected by the context surrounding the features,

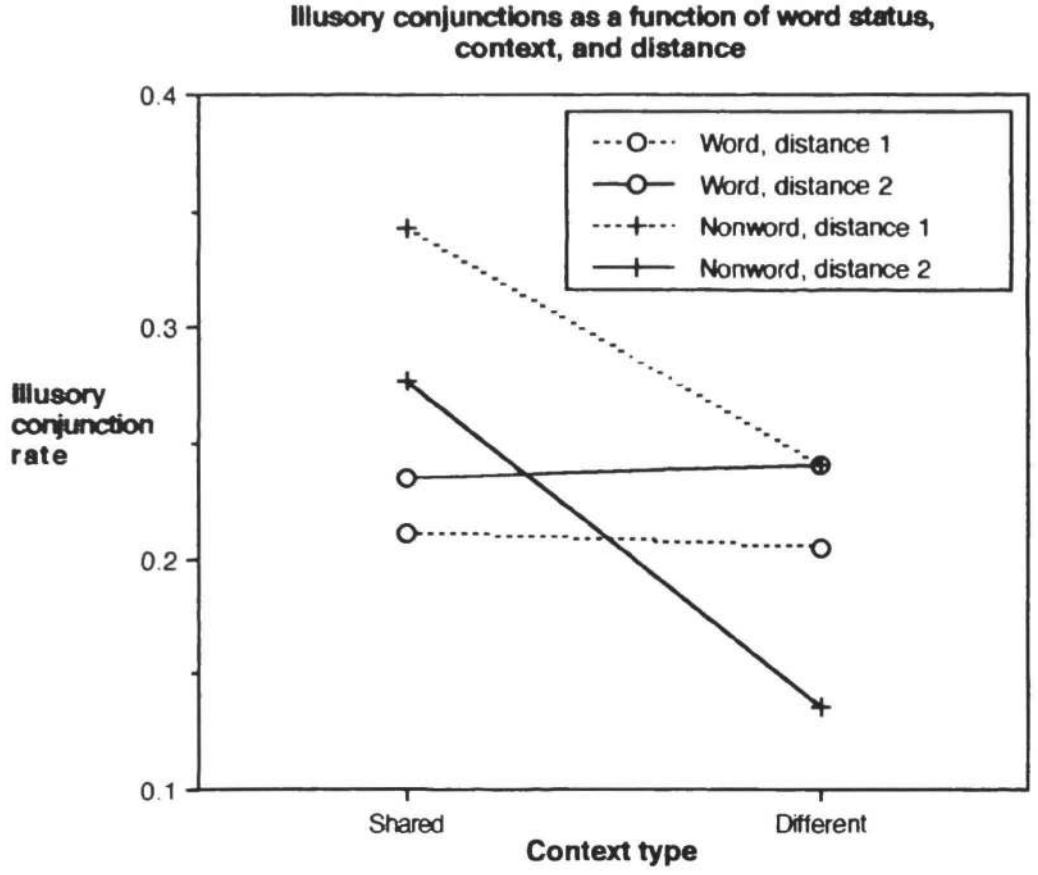

Figure 5. Illusory conjunctions, measured by the false alarm rate for the conjunctive trials.

such that features are more likely to be conjoined if they appear in the same context. This finding suggests that features are encoded in a context sensitive way. But as Figure 5 illustrates, the effect of shared context appears to vary considerably depending upon whether the features appear within a word or nonword. Although the interaction of shared context and word did not quite reach significance, $(1,12) = 3.85$, $p = 0.073$, the effect of shared context can be clearly seen for the nonwords, but not at all for the word stimuli.

The effect of distance was also modulated by the word nature of the stimuli. There was a significant interaction between the word and distance factors, $F(1,12) = 8.18$, $p = 0.014$, so that the effects of distance were more pronounced in nonwords than in words. The principle of encoding similarity appeared to hold in the nonwords, in which adjacent features were more likely to be conjoined than nonadjacent features. If there is any effect of distance on word items, it is that adjacent features are less likely to be conjoined than nonadjacent ones. However, the magnitude of the distance effect is much smaller for words than nonwords, 3.0% versus 8.5%.

Despite the influence of word displays on other experimental variables, there was no evidence of any main effect of word versus nonword displays, $F(1,12) = 0.416$. This result may at first seem at odds with the findings of Prinzmetal and Millis-Wright (1984), who reported fewer conjunctive errors in nonword displays than word displays. In the current experiment, most conditions showed fewer errors for word than nonword stimuli; however, when conditions were optimized for error-free performance based on the principle of encoding similarity (different context, distance 2), there were fewer errors on the nonword than the word items. The current finding therefore offer some limits to the generality of the Prinzmetal and Millis-Wright results.

Discussion

The pattern emerging from these results suggests that the representation of a feature depends both upon local attributes of the feature, like its location and immediate context, as well as more global properties of the feature

environment, such as whether the feature appears within a word or nonword. In the current experiment, the representation of features in nonwords included information about feature location and aspects of the immediate context. At least for nonwords, this location and contextual information is used in the feature binding process, as evidenced by the greater number of illusory conjunctions occurring between features sharing similar locations and contexts.

But another sort of feature representation seems to be established when features appear within words. The current results suggest that the final representation of features can be determined at least in part by knowledge about higher order structures, like words. However, the influence of word structure on feature representation is a contextual effect very different from the effect of the flanking horizontal and vertical contexts. It appears that the feature representation established as a result of the interaction with word knowledge is not sensitive to the specifics of feature instances, like location or surrounding context. Instead, it seems that these local attributes of features are overridden by the influence of higher level organization. This raises the interesting question of how feature binding might proceed in cases where top-down influences are relevant. An interesting sidelight is that although the word factor interacted with other experimental factors, a post-experiment survey found only one subject who was aware of any words being presented during the experiment.

The most interesting result of the experiment may be the evidence it provides for context sensitive feature encoding, since it appears that the feature binding processes use both location and contextual information to bind features together. However, implementing a contextually encoded feature representation raises a number of difficult computational problems. One concern is the possible size of the feature encoding space. The horizontal and vertical contexts used here are one simple example of a context that may be encoded at the feature level, but there is a potentially enormous space of possible contextual encodings. Coarse coded representations are one way of reducing the number of detectors required to span a given encoding space, and Hinton (1981) has shown that the savings advantage of coarse coding increases with the number of encoded dimensions. Other savings could be introduced by using dynamically programmed feature detectors. A relatively small set of detectors might be programmed to encode task relevant context information. A purely speculative approach at this point might be an implementation based upon the CID model of McClelland (1985).

A second problem in implementing contextual maps is using information from a nonspatial contextual map in conjunction with feature maps based on other types of spatial and nonspatial metrics. Two spatially organized maps for different features can be meaningfully compared and correlated, since a direct correspondence between locations in the two maps can be established. However, different contextual maps will be laid out according to different metrics, and there may be no direct mapping from positions within one type of context map to those in another. One way around this difficulty may be to link feature maps based upon different metrics indirectly, through a higher, object level representation. In this scheme, each map would act as a source of constraint on the number and description of objects in the environment. The final representation of the environment would be determined by a constraint satisfaction process using both location and contextual information, operating over the entire set of feature maps. In the resulting representation, each object in the environment description would be linked to the relevant values on all the feature maps. In this way, corresponding positions on differently structured maps could be found by specifying a particular object.

References

Cohen, A. & Ivry, R. (1989). Illusory conjunctions inside and outside the focus of attention. Journal of Experimental Psychology: Human Perception and Performance, 15, 650-663.

Duncan, J. (1989). Parallel processing: Giving up without a fight. Commentary. *Behavioral and Brain Sciences, 12,* 402-403.

Hinton, G.E. (1981) Shape representation in parallel systems. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 1088-1096.

McClelland, J.L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science, 9,* 113-146.

Prinzmetal, W. & Millis-Wright, M. (1984). Cognitive and linguistic factors affect visual feature integration. *Cognitive Psychology, 16,* 305-340.

Strong, G.W., & Whitehead, B.A. (1989). A solution to the tag-assignment problem in neural networks. *Behavioral and Brain Sciences, 12,* 381-433.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Treisman, A. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107-141.

Wolfe, J.M., Cave, K.R., & Franzel, S.L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 419-443.