Published in final edited form as:

Vis cogn. 2011 April 1; 19(4): 469–482. doi:10.1080/13506285.2011.552894.

Visual Learning of Statistical Relations Among Non-adjacent Features: Evidence for Structural Encoding

Elan Barenholtz^{*} and

Department of Psychology, Florida Atlantic University, Boca Raton, FL

Michael J. Tarr

Center for the Neural Basis of Cognition / Department of Psychology, Carnegie Mellon University, Pittsburgh, PA

Abstract

Recent results suggest that observers can learn, unsupervised, the co-occurrence of independent shape features in viewed patterns (e.g., Fiser & Aslin, 2001). A critical question with regard to these findings is whether learning is driven by a structural, rule-based encoding of spatial relations between distinct features or by a pictorial, template-like encoding, in which spatial configurations of features are embedded in a ‘holistic’ fashion. In two experiments, we test whether observers can learn combinations of features when the paired features are separated by an intervening spatial ‘gap’, in which other, unrelated features can appear. This manipulation both increases task difficulty and makes it less likely that the feature-combinations are encoded simply as larger unitary features. Observers exhibited learning consistent with earlier studies, suggesting that unsupervised learning of compositional structure is based on the explicit encoding of spatial relations between separable visual features. More generally, these results provide support for compositional structure in visual representation.

How are the spatial relations between individual visual features learned from complex visual input, such as natural scenes, for later recognition? A number of studies have demonstrated that observers can learn statistical contingencies among disparate visual features in unsupervised tasks (Fiser & Aslin, 2001, 2005; Orban, Fiser, Aslin, & Lengyel, 2008). In such experiments, observers viewed patterns made of multiple shape features arranged in a grid. Unknown to the observers at the initiation of the experiment, the features making up the grid patterns consisted of ‘base’ pairs or triples of features that always appeared together in the same spatial configuration. Observers in these experiments showed an ability to identify the base combinations as belonging together in subsequent testing.

A central question that applies to such studies is whether learning the relations between features is driven by a holistic, ‘pictorial’ form of learning or a rule-based, ‘structural’ form of learning, a dichotomy at the root of different theoretical approaches to object recognition. In a holistic representation, encoding and recognition proceed on the basis of comparing some presented pattern to an image ‘whole-cloth’ without extracting features or their

^{*} Corresponding Author: 777 Glades Road, Boca Raton FL elan.Barenholtz@fau.edu.

relations—essentially a template. This approach is represented in template-based models of visual recognition such as “Recognition-By-Alignment” (Ullman, 1989), in which a stored representation is matched to a new example on the basis of a global transformation of the entire image. The alternative, ‘structural’ approach is based on a representation in which independent features and relations are explicitly encoded. According to this approach, recognition proceeds on the basis of first identifying a set of features and then determining whether they are in the appropriate relations to one another. An example of this form of recognition is reading, in which the evidence suggests that individual letters must first be identified in order for the word that they comprise to be recognized (Pelli, Farell, & Moore, 2003). This latter type of approach is represented by a number of influential models of object recognition (e.g., Biederman, 1987; Marr & Nishihara, 1978) in which features and their relations are both represented explicitly¹.

All of the previously mentioned studies of statistical visual pattern learning in static displays (e.g., Fiser & Aslin, 2001) are consistent with either of the approaches to recognition discussed above². More specifically, subjects in these earlier studies could be learning a structural ‘rule’ that feature X and feature Y always co-occur in a specific spatial configuration, or they could be learning an image template that encompasses both feature X and feature Y—as well as the intervening space between them. In particular, the latter is a plausible explanation for these earlier studies because the paired features were always spatially adjacent to one another. Thus, even though the patterns appear to be made up of different features separated by white space and/or grid lines, it is possible, and perhaps likely, that the repetition of the exact same image (e.g., the image containing feature X and feature Y in a specific configuration) leads to an encoding of the entire pattern as a whole, *aka* a template. Indeed, one result described by Fiser and Aslin (2005), seems to potentially favor a holistic account. In that study, pairs of feature that appeared as part of larger combinations (such as triplets) were not learned as well as pairs that did not appear as a subset of a larger configuration. This seems to be more readily explained by a holistic interpretation, in which the repeated triplet patterns were encoded as a unitary image and that a sub-portion of this pattern did not register a match. A structural account would have more difficulty explaining these results since it is not possible to encode the relation among three individual features without also encoding the relations between pairs of features making up the triplet.

In order to more specifically address whether structural encoding takes place when learning feature relations from complex patterns, our study introduced a learning task similar to Fiser and Aslin's, but one in which members of base-feature pairs were spatially non-adjacent — they were separated by an intervening ‘gap’ in which other, unrelated features appeared. The introduction of an intervening gap itself makes the task more difficult, simply because the to-be-learned relations occur across a larger visual distance. More importantly, we reasoned that the presence of an intervening space encompassing variable features between consistent features makes it unlikely that the base pairs will be encoded as single ‘holistic’ features,

¹By “explicitly” we mean within the representation; no reference to the notion of “explicit” memory is intended here.

²Indeed, as discussed below, this question applies to virtually all studies of encoding relations between distinct features (see Barenholtz and Tarr, 2007; Pothose, 2005).

that is, encoding a single image that is repeated consistently (as in earlier studies). Instead, learning feature pairs is presumed to require a structural encoding based on the spatial relationship between distinct features.

Experiment 1

Methods

Subjects—15 Brown University undergraduate students participated in exchange for cash payment.

Stimuli and Procedure—Each subject performed two separate blocks of trials: a Learning Phase and a Test Phase.

Stimuli in the Learning Phase consisted of polygonal shapes (“features”) arranged in a 4×4 grid of possible locations (Figure 1). Unlike the studies discussed earlier, in our study the grid was not visible. Each of the individual shapes measured approximately 1.3×1.3 degrees of visual angle and consisted of a bilaterally symmetrical polygon with a distinct color (color was included in our experiments in order to make the task easier in an attempt to offset the greater difficulty inherent in learning relations among spatially distant features). Of the 22 shapes available, a subset of 16 were chosen and assigned to one of three categories: *Vertical pairs* (6 shapes comprising 3 pairs), *Horizontal Pairs* (6 shapes comprising 3 pairs), and *Singletons* (4 shapes). Vertical pairs always appeared together within a pattern and were always in a configuration in which one of the features from the pair (the same one on every trial) was two grid spaces above the other, horizontally aligned, within the grid. Similarly, the Horizontal Pairs always appeared together in a configuration in which one of the features was two grid spaces to the right of the other, vertically aligned, within the grid. On any given trial, either two or three pairs appeared in the overall pattern, consisting of either one Horizontal Pair and one Vertical Pair, two Horizontal Pairs and one Vertical Pair or two Vertical Pairs and one Horizontal Pair. The specific choice of pairs was randomized across trials. Singletons did not have any specific relation to any of the individual paired shapes but appeared within the space between the vertical or horizontal pairs on any trial when that space was otherwise empty (i.e., when no member of the other pairs appeared within that space). This ensured that the base pairs were always separated by another, unrelated, feature.

During the Learning Phase, the experimental sequence was as follows: On each trial a single pattern appeared in a random location on the screen within 4 degrees horizontally and vertically offset from the center of the screen. After two seconds, the pattern disappeared, followed by a gray screen for 1 second. Then a second pattern appeared in a new random location for one second, after which it disappeared. This change in the location of the grid was intended to eliminate a relation between the absolute screen position and the shape features; that is, encouraging subjects to learn the relative positions of the features in the grid. On one-half of the trials (‘No-change Trials’), the second pattern was identical to the first one while in the other half (‘Change Trials’), either one of the base pairs or one of the singletons was switched for another pair or singleton that had not been shown in the first pattern. To ensure that subjects needed to attend to the individual features, and not just the

global shape of the pattern, the new pair or singleton introduced on Change trials was always in the same location as the one it replaced. Note that the change never resulted in a member of a base pair appearing with some other feature in the paired position other than its respective pair.

In the Test Phase, subjects discriminated between true base pairs from the Learning Phase and ‘non-pairs’: pairs of features selected from two different base pairs that were conjoined in either a horizontal or vertical orientation for the Test Phase only. There were a total of six non-pairs— three horizontally configured and three vertically configured—to match the six true base-pairs. The test sequence was as follows: On each trial, subjects saw a single pair of features—consisting of two features shown in either a horizontal or vertical configuration, without a singleton between them—for two seconds. This was followed by a gray screen for one second and then another pair of features for two seconds. On one-half of the trials, the true base pair was shown first, followed by the non-pair and on the other half of the trials this order was reversed. After the two pairs had been shown, subject were asked to determine which of the two displayed pairs, the first or the second, looked more familiar (based on the patterns they had viewed in the Learning Phase). Each subject saw the six base pairs compared to each of the six non-pairs for a total of 36 trials per subject.

Results and Discussion

The overall accuracy in the Learning Phase was 78%. In the Test Phase, we calculated the percentage of trials on which subjects chose, on average, the Base Pair as being more familiar than the Non-Pair. There was a tendency to choose the Base pairs as being more familiar than the Non-Pairs (Mean=62%; S.E.= 2%; range 47% to 72%). A *t*-test comparing observed performance to hypothetical chance performance found that this tendency was significant ($t=3.267$ $p<.001$). This level of performance is comparable to earlier studies of statistical learning (e.g., Experiment 2 of Fiser and Aslin, 2001). On the one hand, this is somewhat surprising since the task in our study would appear to be much more difficult as it requires that subjects encode relations among spatially distant features. One possible advantage in our experiment however, compared with these earlier studies, is that the features contained color information, whereas earlier studies used black and white stimuli making any direct comparison between performance levels difficult.

In post-experiment interviews, we asked subjects whether they had noticed that certain features appeared in pairs during the Learning Phase. Only one subject reported explicitly noticing the paired stimuli while most subjects reported being unaware of any statistical structure during the Learning phase. These results are consistent with recent reports concerning the ‘implicit’ nature of statistical learning (Kim, Seitz, Feenstra, & Shams, 2009).

Experiment 2

Because the members of a base pair always appeared together and always in the same configuration, there are two possible sources of information that subjects might have used to distinguish between Base-Pairs and Non-Pairs in the Test task: the constant conjunction of the paired features (i.e., A always appears with B) and their configuration (i.e., A always

appears to the right of B). To determine whether subjects encoded configural information, and not just the constant conjunction, Experiment 2 had subjects perform the same Learning task but with a Test task in which they had to choose between learned and novel configurations of the same features.

Methods

10 Brown University undergraduates who did not participate in Experiment 1 and naïve to the purpose of the study participated for class credit.

Stimuli and Procedure

The stimuli and procedure were identical to Experiment 1 except for the nature of the comparison task in the Test Phase. On each trial of the Test Phase in Experiment 2, subjects had to choose between base pairs and ‘reverse pairs’, consisting of the same features as the base pair but in reverse configuration (i.e., if the base pair consisted of A shown to the right of B during learning, the reverse pair consisted of A shown to the left of B). As in Experiment 1, each of the true base pairs was compared with each of the reverse base pairs for a total of 36 trials per subject.

Results and Discussion

The overall accuracy in the Learning Phase was 82%. In the Test Phase there was a tendency for subjects to choose the base-pairs as more familiar than the reverse pairs (Mean=66%, SE=.03; range 41% to 78%). A *t*-test comparing observed performance to hypothetical chance performance found that this tendency was significant $t(10)=5.29$, $p<.0001$. These results demonstrate that subjects learned the spatial configuration of the members of the pairs, not just their constant conjunction.

General Discussion

Our study extends earlier experiments showing unsupervised learning of relations between distinct visual features to cases in which the related features are spatially non-adjacent. Unlike earlier studies, in our experiments the learned pairs did not constitute a single, consistently repeated image, which could serve as a template to be matched directly to a test stimulus. This makes a holistic ‘pictorial’ representation much less likely, although not impossible. For example, one possible way of doing the task using a simple template-based system would be to determine that the correct test pair is *more* globally similar—though not identical—to the images presented during learning. However, without highlighting the stable pairs, there is no means for establishing which images are ‘important’ during learning, effectively requiring that *all* of the presented patterns be stored equally. A more plausible interpretation is that learning depends on somehow detecting the stable relations among features and encoding that information while ignoring the arbitrarily intervening features. Such an encoding scheme, which depends on detecting relations among spatially distant features, is inherently structural.

Structural models offer a number of theoretical advantages over alternative approaches. As in feature-based models, they can provide representational flexibility for generalizing across different members of a category, however, they also provide additional constraints to avoid

overgeneralization. For example, identifying a ‘smiley face’ on the basis of a collection of features (e.g., two circles and a curved arc), rather than an entire image of a smiley face, allows identification across examples in which the spacing between the features varies. Of course, the potential downside is that such features may be highly generic and therefore *underconstrain* the model, leading to false positive matches. The alternative—to use more specific features, such as a *pair* of circles (the ‘eyes’ of the smiley face)—may *overconstrain* the model, since it would then depend on a specific configuration of features whose spacing actually varies across examples. Encoding relations along with the features, that is, structure, can serve to overcome some of these problems. As long as the relations are defined abstractly (e.g., ‘above’, ‘horizontal to’), a wide degree of flexibility is preserved while providing critical constraints to reduce false positive matches for otherwise generic features.

The potential utility of structural encoding has long been recognized and several influential theories of visual object recognition have been explicitly structural (Biederman, 1987). On the empirical side, a number of experimental studies have looked directly at the role of relations in the encoding of complex objects (Hummel & Stankiewicz, 1997; Saiki & Hummel, 1998). However, as noted, experimentally isolating the role of configurations between features is often difficult since any configuration of features may, in theory, be encoded on the basis of a larger, ‘holistic’ feature that encompasses the smaller ones (for an extensive discussion of this issue, see Barenholtz & Tarr, 2007). For example, Arguin and Saumier (2004) compared visual search times for when the search target and distracters shared either features (volumetric parts) or configurations (the spatial organization of the parts). They found that these two properties—parts and relations—contributed independently to search times. Similarly, (Keane, Hayward, & Burke, 2003) used a change detection task to assess the perceptual similarity of objects that had shared parts in different configurations versus shared configurations of different parts, finding that shared configurations yielded higher perceptual similarity (i.e., worse change detection). However, in both of these studies, objects that shared configurations of different parts also shared similar global shapes. Thus, it is possible that subjects in these experiments did not explicitly encode the configural information between separate parts *per se*. A similar line of reasoning applies to a number of studies of face recognition that aim to compare the role of individual features of a face—such as the eyes and mouth— and their configuration—such as the spacing between the eyes (Maurer, Le Grand, & Mondloch, 2002; Tanaka & Farah, 1993). Here too, differences in configuration may be captured by using larger, ‘holistic’ facial features rather than explicitly encoding configural information (Zhang & Cottrell, 2005; Nestor, Vettel, & Tarr, 2008). The current study, while providing perhaps some of the strongest evidence to date for structural encoding, does not explicitly address the role of structure in typical applies recognition tasks such as face and object recognition. Thus, further studies will be needed to address the scope and application of this type of learning.

It is worth noting that while studies of *visual pattern* learning prior to ours have used *spatially* adjacent features, a number of studies of statistical learning in the *auditory* domain have used *temporally* non-adjacent features. These studies demonstrate that people—as well as non-human primates—can learn the temporal relations between pairs of non-adjacent

features (Creel, Newport, & Aslin, 2004; Gebhart, Newport, & Aslin, 2009; Newport, Hauser, Spaepen, & Aslin, 2004). In these studies, subjects were able to learn the relations among paired sounds even when the paired sounds were separated by an intervening, unrelated sound. Such findings suggest interesting parallels to the results of our present study. However it is impossible to say at this point whether these superficially similar results reflect a more general learning mechanism concerned with relations between distinct features – regardless of the perceptual dimension – or whether they reflect distinct mechanisms particular to each different dimension (in this case time and space). In particular, learning sequences of auditory signals depends on cognitive systems devoted to temporal encoding, even perhaps representing specific linguistic or pre-linguistic mechanisms. In contrast, visual pattern learning of the sort described here depends on cognitive systems devoted to spatial encoding and visual recognition. Thus, there may be little overlap in the underlying mechanisms supporting the learning of non-adjacent features in different perceptual domains. At the same time, the possibility that such learning depends on a more general, shared statistical mechanisms is an intriguing one that merits further research.

Acknowledgments

This research was supported in part by NGA Award #HM1582-04-C-0051 and an NIH EUREKA Award #1R01MH084195-01.

References

- Arguin M, Saumier D. Independent processing of parts and of their spatial organization in complex visual objects. *Psychological Science*. 2004; 15(9):629–633. [PubMed: 15327635]
- Barenholtz, E.; Tarr, MJ. Reconsidering the role of structure in vision.. In: Markman, AB.; Ross, BH., editors. *Categories in use. The psychology of learning and motivation: Advances in research and theory*. Academic Press; New York: 2007.
- Biederman I. Recognition-by-Components - a Theory of Human Image Understanding. *Psychological Review*. 1987; 94(2):115–147. [PubMed: 3575582]
- Creel SC, Newport EL, Aslin RN. Distant Melodies: Statistical Learning of Nonadjacent Dependencies in Tone Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2004; 30(5):1119–1130.
- Fiser, J. z.; Aslin, RN. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*. 2001; 12(6):499–504. [PubMed: 11760138]
- Fiser, J. z.; Aslin, RN. Encoding Multi-element Scenes: Statistical Learning of Visual Feature Hierarchies. *Journal of Experimental Psychology: General*. 2005; 134(4):521–537. [PubMed: 16316289]
- Gebhart AL, Newport EL, Aslin RN. Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*. 2009; 16(3):486–490. [PubMed: 19451373]
- Keane SK, Hayward WG, Burke D. Detection of three types of changes to novel objects. *Visual Cognition*. 2003; 10(1):101–127.
- Kim R, Seitz A, Feenstra H, Shams L. Testing assumptions of statistical learning: is it long-term and implicit? *Neuroscience Letters*. 2009; 461:145–149. [PubMed: 19539701]
- Marr D, Nishihara HK. Representation and Recognition of Spatial-Organization of 3-Dimensional Shapes. *Proceedings of the Royal Society of London Series B-Biological Sciences*. 1978; 200(1140):269–294.
- Maurer D, Le Grand R, Mondloch CJ. The many faces of configural processing. *Trends in Cognitive Sciences*. 2002; 6(6):255–260. [PubMed: 12039607]

- Nestor A, Vettel JM, Tarr MJ. Task-Specific codes for face recognition: How they shape the neural representation of features for detection and individuation. *PLoS ONE*. 2008; 3(12)
- Newport EL, Hauser MD, Spaepen G, Aslin RN. Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*. 2004; 49(2):85–117. [PubMed: 15304368]
- Orban G, Fiser J, Aslin RN, Lengyel M. Bayesian learning of visual chunks by human observers. *PNAS Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(7):2745–2750.
- Pelli DG, Farell B, Moore DC. The remarkable inefficiency of word recognition. *Nature*. 2003; 423(6941):752–756. [PubMed: 12802334]
- Pothos EM. The rules versus similarity distinction. *Behavioral and Brain Sciences*. 2005; 28(1):1–49. [PubMed: 16047456]
- Saiki J, Hummel JE. Connectedness and part-relation integration in shape category learning. *Memory and Cognition*. 1998; 26:1138–1156. [PubMed: 9847542]
- Tanaka JW, Farah MJ. Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology*. 1993; 46A:225–245. [PubMed: 8316637]
- Ullman S. Aligning pictorial descriptions: An approach to object recognition. *Cognition*. 1989; 32(3): 193–254. [PubMed: 2752709]
- Zhang, L.; Cottrell, GW. Holistic processing develops because it is good.. In: Bara, BG.; Barsalou, L.; Bucciarelli, M., editors. *Proceedings of the 27th Annual Cognitive Science Conference*; Mahwah: Lawrence Erlbaum. 2005.

Figure 1.

Example stimuli used in Experiments 1 and 2. **Top:** The stimuli consisted of patterns of Vertical and Horizontal ‘base-pairs’, which always appeared together in the same configuration, as well as Singletons which appeared in the gap between base-pair members whenever that location was not occupied by a feature from another pair. **Bottom:** Two example stimulus patterns made up of the base pairs and singletons shown above.