6

Perceptual Constraints on Implicit Memory for Visual Features: Statistical Learning in Human Infants

Richard N. Aslin

The visual world contains an enormous amount of information that, to a naïve infant, must initially appear to be overwhelmingly complex. Yet within a few weeks after birth, infants are recognizing familiar objects and controlling a variety of motor systems to interact with them. Even if some of these object-recognition mechanisms are triggered by innate biases (e.g., for faces), there are simply too many objects in the visual world to be processed by separate brain mechanisms—especially since the vast majority of objects that we encode in memory are created by our culture. Thus, it is implausible that more than a handful of specialized object-recognition systems could have evolved for our species. We are forced to conclude, therefore, that object recognition must involve a powerful learning mechanism that rapidly encodes into memory both basic-level object categories (e.g., dog, chair) and individual exemplars of those categories (e.g., Fido the terrier, my favorite leather recliner).

A longstanding debate in the object-recognition literature concerns how adults are able to encode thousands of objects, both highly familiar and recently novel, in memory. One view is that objects are encoded as "wholes," or exemplars (Bulthoff, Edelman, & Tarr, 1995), and the other view is that they are encoded at the sub-object level as a collection of "parts" in particular spatial configurations (Biederman, 1987). Evidence in favor of the former view comes from the perception of highly familiar objects, such as faces, whose parts (or features, the eyes/nose/mouth) are poorly recognized when removed from their surrounding context or when presented in a non-canonical orientation (e.g., inverted). Evidence in favor of the latter view comes from observers' generalization across viewpoint, such as recognizing a truck as a truck, regardless of whether it is viewed from the front, the side, the back, or even from above. If objects are not decomposed into their parts, then how could generalization across viewpoint, despite only partial overlap in the features,

This chapter is an expanded version of a talk presented on March 25, 2008, in Vancouver at a festschrift honoring Leslie B. Cohen. The research described in this chapter was supported by grants from NIH (HD-037082) and the McDonnell Foundation (220020096).

be accomplished? Moreover, if the exemplar hypothesis were correct, it would require the storage of an exceptionally large number of objects, because each subtle change in viewpoint would require a different memory representation even for the same object.

One reason that the debate persists between exemplar (wholes) and compositional (parts) theories of object recognition is because judgments made by adults can be partially guided by top-down knowledge. Once a basic-level category (e.g., a cup) has been learned, it is difficult for adults to decompose that object into its constituent parts (e.g., cup = bowl + handle) without conscious effort. One way to circumvent this problem is to study the development of object recognition and category formation in infants, who have little or no top-down knowledge to guide this process. By revealing the fundamental learning mechanisms that capitalize on innate biases, the study of infants has the potential to clarify not only how infants learn about objects, but also how adults are able to develop sophisticated object-recognition skills.

TWO CLASSIC STUDIES OF INFANT CATEGORY-LEARNING

Our understanding of category learning by infants was advanced by two seminal studies from the Cohen lab at the University of Texas in the 1980s. In contrast to research on adult categories, which focused on visual objects from the natural environment (e.g., Smith & Medin, 1981), Younger and Cohen (1983) created novel objects composed of multiple parts in different combinations. As shown in Figure 6.1, each object had five parts (body, tail, feet, ears, legs) and each of these body parts, which we will refer to as visual features, could take on one of three possible values. For example, the tail could be from a bird, a horse, or a rabbit. Thus, the entire inventory of unique objects generated by these feature combinations was 3^5 or 243. From this large inventory there are many ways in which a subset of the 243 objects could be assigned to a category. For example, a category could be defined as "all objects with floppy ears" or "all objects with short legs". Alternatively, a category could be defined by a combination of features, such as "all objects with floppy ears and short legs." It is this latter definition of a category that was tested by Younger and Cohen (1983).

Groups of 4-, 7-, and 10-month-old infants were habituated to four objects selected from the inventory of 243. Two of the four objects shared three features ($body_1$, $tail_1$, $feet_1$) and the other two features were randomly selected. The other two objects also shared the same three features ($body_2$, $tail_2$, $feet_2$), but the specific values of these features were different from those defining the first two objects (again, the fourth and fifth features were randomly selected). Thus, if infants can keep track of feature correlations, they could define the first two objects as belonging to one category and the second two objects as belonging to a second category. After each infant had been familiarized with the four objects (two 20-s trials for each object), they were presented with three post-habituation test objects, all of which were novel. One test object had the same combination of three features (body, tail, feet) that defined one of the

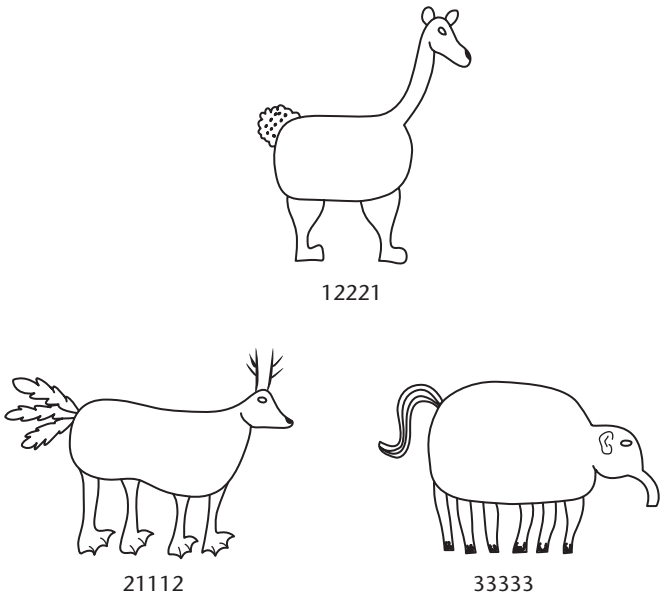

Figure 6.1. Example stimuli from Younger and Cohen (1983) showing the novel objects created by five different features (body, tail, feet, legs, ears), each of which could take on three different values.
Source: From "Infant Perception of Correlations among Attributes," by B. A. Younger and L. B. Cohen, 1983, *Child Development*, *54*, p. 860. Copyright 1983 by Society for Research in Child Development, Inc. Reprinted with permission.

categories, but with a different combination of the two random features. The second test object had a different combination of the three defining features, thereby conforming to neither of the two object categories, even though each feature had been seen during familiarization (e.g., the body from category one and the tail from category two). And the third test object had a completely novel set of the three features, none of which had been seen during familiarization. Thus, if infants can form two categories based on the correlations among the three defining features, and ignore the two random features, they should find the first test trial type (correlated features) to be familiar, the second test trial type (uncorrelated features) to be novel, and of course the third test trial type (novel features) to be novel (this served as a control for fatigue or lack of feature encoding).

The results of Younger and Cohen (1983) were clear. The 10-month-olds showed the pattern of results expected of a learner who has formed object categories: the uncorrelated and novel-feature test objects led to recovery from habituation, whereas the correlated-feature test object (despite having two additional uncorrelated features) showed no recovery. In contrast, neither the 4- nor the 7-month-olds showed recovery to the uncorrelated-feature test object or the correlated-feature test object. This pattern of results suggests that these younger infants were not forming a category based on feature

correlations. Recovery on the novel-feature test rules out the uninteresting possibility that these younger infants did not encode the features. Thus, the overall pattern of results suggests that between seven and ten months of age, infants acquire the ability (at least for this class of objects) to form categories based on feature correlations, and prior to this age they are susceptible to a feature conjunction error (i.e., confusing out-of-category objects that share individual features as if they were members of the category defined by feature conjunctions).

Of course, the specific age at which infants acquire the ability to form categories based on feature correlations could be influenced by a variety of factors, including the number of features, the number of within- and between-category exemplars presented during familiarization, and the ease of discrimination among the features within a set (e.g., variations in leg length or body size). As a first step in exploring whether these task variables might influence the age of onset for category formation based on feature correlations, Younger and Cohen (1986) conducted a series of follow-up experiments using the same basic set of 5-feature objects, but now only the three *relevant* features were varied to define two categories (the other two features were held constant). Thus, in contrast to Younger and Cohen (1983), there were only two rather than four objects presented during the familiarization phase (five 20-s trials for each object). Under these conditions, 7-month-olds showed the pattern of results that the 10-month-olds had shown when all five features varied during familiarization and three relevant features defined the category. That is, 7-month-olds showed recovery to the uncorrelated and novel feature tests but not to the correlated feature test. However, in a second experiment in which two of the three relevant features defined the category, but the two nondefinitional features were held constant, again only the 10-month-olds showed evidence of basing their categories on feature correlations.

In sum, as stated by Younger and Cohen (1983), 10-month-olds "generalized their habituation to a novel test stimulus that maintained the correlation they had seen, whereas they dishabituated to a stimulus containing equally familiar features but that failed to preserve the correlation" (p. 864–865). Seven-month-old infants are capable of using feature correlations to define categories, but only when extraneous features and irrelevant feature correlations are reduced or eliminated. Finally, control conditions confirmed that 4- and 7-month-old infants are attentive to features and encode them, so the fundamental limitation appears to be the mechanism by which features are combined.

FEATURES REVISITED: STATISTICAL LEARNING FROM VISUAL SCENES

The tradition within which the Younger and Cohen (1983, 1986) studies were conducted—how basic mechanisms of information processing influence infant category formation—continues to hold a prominent position in the field (Cohen, Chaput, & Cashon, 2002). However, a different tradition emerged a decade later in the domain of language learning—how temporal-order information

allows 8-month-old infants to group successive sounds into word-like units (Saffran, Aslin, & Newport, 1996). At first blush, these are very different tasks. As instantiated in the studies by Younger and Cohen, infants are allowed to scrutinize a single visual stimulus for up to 20 s at a time and extract, across a series of trials, those features contained in the set of stimuli that are invariant in the face of variation among the other features. In contrast, Saffran et al. (1996) presented infants with a rapid stream of elements (4 speech syllables per second) and asked whether they could extract the statistical coherence (i.e., transitional probability) of successive syllables compared to the relatively incoherent syllable transitions that occur at word boundaries.

Despite these obvious differences, the fundamental task for the infant learner in these two domains is similar: which elements (or features) form clusters based on their co-occurrence relations? In Younger and Cohen's studies, the correlations are defined *within* an image (e.g., long tail and fat body), whereas in Saffran et al., the correlations are defined *across* successive sounds. A substantial body of work from several labs has demonstrated that temporal-order correlations need not be limited to the auditory modality, despite the obvious utility of such a statistical learning mechanism for auditory stimuli. For example, Fiser and Aslin (2002a) and Kirkham, Slemmer, and Johnson (2002) showed that simple visual shapes, presented one at a time in constrained temporal orders, can lead both adults and infants to form temporal-order groupings of these shapes.

More relevant to the Younger and Cohen studies is whether a statistical learning mechanism that is sensitive to the coherence relations among spatially distributed object parts, elements, or features, can account for visual category formation. A first step in answering that question is the demonstration that adults can, by mere exposure to a series of visual images, learn which elements "go together" to form feature conjunctions. Fiser and Aslin (2001) created a set of 144 scenes, each of which consisted of six simple shapes arranged in a cluster (i.e., each shape was immediately adjacent to at least one other shape). Unbeknownst to the participants, the shapes in each scene came from an inventory of 12 shapes that were organized spatially into pairs (2 horizontally arranged shape-pairs, 2 vertically arranged shape-pairs, and 2 obliquely arranged shape-pairs). Shape-pairs were then randomly selected from the inventory, and three of them were clustered to create the set of 144 scenes. The key question is whether the pairs of shapes, because of their *spatial* coherence across the set of scenes, would become recognized as "units of perception" compared to other pairs of shapes that did not consistently appear in spatial proximity across the set of scenes. The answer was a clear yes—after only 5–20 minutes of passive observation, in which each of the 144 scenes was only visible for two seconds, adults judged a spatially coherent pair of shapes as more familiar than a pair of shapes that appeared less consistently in the scenes.

Fiser and Aslin (2002b) extended this work with adults to 9-month-olds by creating a set of 16 scenes, each of which was composed of three simple shapes – a pair of shapes that always appeared in the same spatial arrangement, and a third shape that appeared adjacent to the shape-pair but in different spatial

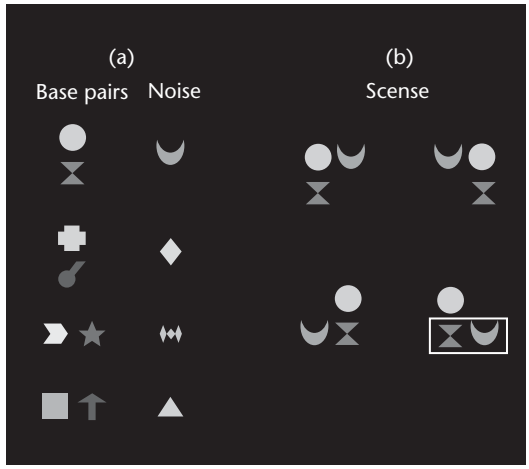

Figure 6.2. Example stimuli from Fiser and Aslin (2002b) showing (a) sample base-pairs of shapes and the third "noise" shape that created less coherent statistics with the other shapes when combined in (b) 3-shape scenes. The white rectangle shows a pair of less coherent shapes (one base-pair element + a noise element) that was contrasted with a coherent base-pair during the post-habituation test phase. (See also figure in plate section.)
Source: From "Statistical Learning of New Visual Feature Combinations by Infants," by J. Fiser and R. N. Aslin, 2002, *Proceedings of the National Academy of Sciences*, *99*, p. 15823. Copyright 2002 by National Academy of Sciences, U.S.A.

arrangements (see Figure 6.2). As in the adult study, the key question was whether 9-month-olds would learn by mere observation that some shape-pairs appear consistently in specific spatial arrangements, whereas other shape-pairs do not. The standard infant-controlled habituation paradigm was used to expose 9-month-olds to repeated 2-s presentations of the 16 scenes (each containing 3 shapes), and in a posthabituation test phase each infant viewed statistically coherent and less coherent shape-pairs. The results were clear in showing that infants look longer to coherent shape-pairs than to less coherent shape-pairs. Moreover, a follow-up experiment showed that they learn the statistical coherence of shape-pairs by extracting either of two distributional properties across the 16 scenes: (a) the relative frequency of shape-pairs (i.e., how many times each shape-pair appears during the habituation phase), and (b) the conditional probability of shape-pairs (i.e., the likelihood that each shape co-occurs with every other shape, regardless of overall shape frequency).

The results of Fiser and Aslin (2002b) not only demonstrate that infants, like adults, can extract by passive observation the statistical coherence of elements in multi-element scenes, but also suggest that statistically coherent elements serve as "parts" that can be recombined to form new objects. As stated by Fiser and Aslin: "Infants preferentially attend to the previously extracted features when they are subsequently presented 'out of context' in different displays." Younger and Cohen (1986) made a similar point in their

studies of category formation via feature correlations: "the bulk of the evidence we have presented is consistent with the notion of a developmental trend from processing parts to processing wholes to processing invariant feature combinations" (p. 814).

Despite the similarities between Younger and Cohen (1983, 1986) and Fiser and Aslin (2002b), there are a number of differences. First, Younger and Cohen asked whether infants could learn the similarity of spatially *nonadjacent* elements (e.g., tail and ears) while ignoring extraneous and often spatially adjacent elements that were not relevant to the underlying category. For example, in Younger and Cohen (1983) out of 40 feature-pairs presented during the habituation phase, only 12 were spatially adjacent, and only one (body-tail) covaried during habituation (to define the category). In contrast, in Fiser and Aslin, all of the relevant statistical relations among elements involved the coherence of spatially *adjacent* elements; the statistical relations among nonadjacent elements were uniformly low and were never tested.

A second fundamental difference between Younger and Cohen (1983, 1986) and Fiser and Aslin (2002b) is the form of the test stimuli. Younger and Cohen used whole objects that consisted of all five features, although only a small subset of the feature combinations was tested. In contrast, Fiser and Aslin used parts of multi-element scenes (or feature combinations) as the test stimuli. Thus, the Younger and Cohen test was more difficult in the sense that infants had to search the entire test scene to determine whether a feature correlation was present (or not), whereas the Fiser and Aslin test extracted the critical information (element coherence) from the larger scene, thereby eliminating the search process. To paraphrase what an adult might conclude about these two tests, Younger and Cohen assessed whether there is something "wrong" in the cluttered scene (e.g., a feature that is out of place), whereas Fiser and Aslin assessed whether part of the scene makes sense (e.g., an entire face versus part of a face covered by the shadow of a tree branch). As noted by Younger and Cohen (1986): "it is not clear whether 10-month-old infants encoded the correlated parts as a single unit or feature, or whether they perceived an invariant relation among separate parts" (p. 814).

CONSTRAINTS ON VISUAL STATISTICAL LEARNING

As noted above, the fundamental differences between the correlated feature approach adopted by Younger and Cohen (1983, 1986) and the statistical learning approach adopted by Fiser and Aslin (2002b) are subtle. Both had a goal of characterizing whether infants can learn correlations among elements in visual scenes to define objects (and their parts) and to use these correlations as defining features for object categories. Younger and Cohen began the process of mapping the correlational landscape to define the information processing constraints that enable category learning. Fiser and Aslin provided a quantitative metric (element co-occurrence and conditional probability) that could be used to characterize which correlations are learned and which are treated as noise.

A key organizing principle in the statistical learning approach is that even simple objects and scenes contain a large number of *potentially relevant* correlations. This, of course, was one of the design parameters manipulated by Younger and Cohen (1983, 1986). But correlations are only one of a large family of potentially relevant *statistics*. Fiser and Aslin (2002b) studied two such statistics that could enable infants to extract shape-pairs from simple scenes (relative frequency and conditional probability). There are many other statistics (e.g., single-element frequency, shape-triplet frequency, conditional entropy, and a variety of higher-order statistics that combine spatial and temporal correlations across scenes). The task facing the infant learner is to efficiently deploy their limited information processing capacity so that they extract the *relevant* statistics and are not overwhelmed by all the available but irrelevant statistics in the set of images to which they are exposed.

There are several ways in which the problem of "too many statistics" could be solved (or at least reduced). First, there are a variety of Gestalt principles by which elements within scenes are linked together by "automatic" mechanisms. For example, elements cohere when they move together (common motion), when they share a common orientation (good continuation), and when they have the same shape (grouping by similarity). Each of these Gestalt principles has been demonstrated in young infants (common motion in Kellman & Spelke, 1983; good continuation in Gerhardstein, Kovacs, Ditre, & Feher, 2004; similarity grouping in Quinn & Bhatt, 2005). Figure 6.3a shows an embedded figure (a circle) within a cluttered array of line segments; both adults and 3-month-old infants can extract the circle almost immediately (without the need for feature learning).

Although it is commonly accepted that Gestalt cues are innate (i.e., provided to organisms by an evolutionary process based on the statistics of natural scenes), there is evidence that some Gestalt cues are not effective in early infancy (cf., Quinn & Bhatt, 2005). Thus, it is possible that a statistical learning

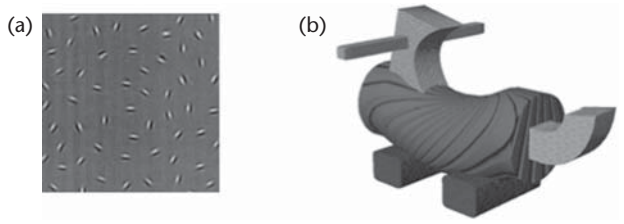

Figure 6.3. Sample images illustrating (a) how the Gestalt principle of good continuation enables the immediate perception of connected contours (the central circular form) in a cluttered array of contours. (b) how a complex, multi-part object presents a challenge for learning which features define a part. (See also figure in plate section.)
Source: From "Detection of Contour Continuity and Closure in Three-Month-Olds," by P. Gerhardstein, I. Kovacs, J. Ditre and A. Feher, 2004, *Vision Research*, *44*, p. 2982. Copyright 2004 by Elsevier. Reprinted with permission. Images reprinted from Michael J. Tarr, Brown University, http://www.tarrlab.org/.

mechanism induces some Gestalt principles via postnatal visual experience, thereby failing to solve the problem of an exponentially large number of potential statistics (at least in the immediate postnatal period). However, if the natural visual environment has a hierarchy of visual statistics, which we know to be true from the work of Geisler (2008), then perhaps the most common statistics are extracted first, and serve as a set of constraints for the extraction of less common statistics. This would be an example of statistical bootstrapping.

A second way in which the number of potentially relevant statistics could be reduced is via an intrinsic hierarchy of stimulus salience. For example, we know that even newborns have preferences for certain types of stimuli, such as high-contrast edges. If attention is directed primarily to a limited subset of visual stimuli, despite a much larger set of elements in the visual field, then this in turn reduces the set of potentially relevant statistics. Unfortunately, it is difficult to define stimulus salience except for simple dimensions (e.g., contrast) that can be varied quantitatively to determine a threshold for detection. For example, it is not clear in the stimuli used by Younger and Cohen (1983; see Figure 6.1) whether all of the five features were equally salient to the infants at 4, 7, and 10 months of age, despite the fact that infants were sensitive to a change in a single feature. That is, the relative salience of features is likely to be affected by the context in which they are presented, rendering any estimate of salience to be nonlinear. An example of how context can affect feature salience is illustrated by comparing Figures 6.3a and 6.3b – the salience of a given line segment is influenced by the surrounding features.

Typically, experiments utilize counterbalanced designs to "average out" any salience effects. However, two recent studies provide compelling evidence that this design strategy is unlikely to be sufficient. Civan, Teller, and Palmer (2005) showed that post-familiarization preferences for paired visual stimuli are influenced by the magnitude of pre-exposure preferences. A highly preferred stimulus is unlikely to lead to a post-familiarization novelty preference, whereas a dispreferred stimulus almost always leads to a post-familiarization novelty preference. Thus, if pre-familiarization preference is not taken into account, negative evidence of post-familiarization preference could be incorrectly interpreted as a failure of discrimination.

Kaldy, Blaser, and Leslie (2006) asked whether 6.5-month-old infants could remember differences in one stimulus dimension (color) better than differences in a second stimulus dimension (luminance). To eliminate the likely possibility that any differences in memory performance were due to dimensional salience, they conducted a separate study to select only equisalient stimuli that varied along both dimensions in a subsequent test of memory. As shown in Figure 6.4, the red standard stimulus was matched with a yellow comparison stimulus so that the color and luminance differences were equally salient (as determined in a preference test). In this way, they could assess memory differences for color and luminance that were unconfounded by salience. Such a design is a model for studies of statistical learning because it renders any extraction of element co-occurrences independent of a salience hierarchy.

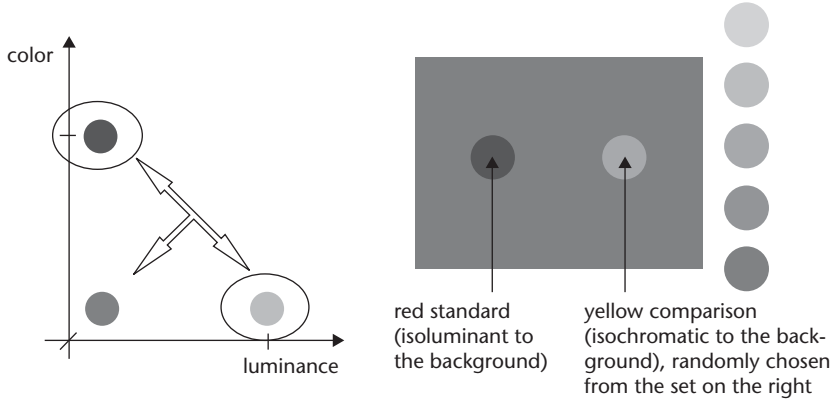

Figure 6.4. A two-dimensional stimulus space (color × luminance) illustrating how a standard stimulus (circled in the left plot) can be rendered equisalient along both dimensions when a comparison stimulus (in the right plot) is varied along one dimension (luminance) and paired with the standard stimulus. (See also figure in plate section.)
Source: From "A new method for calibrating perceptual salience across dimensions in infants: The case of color vs. luminance," by Z. Kaldy, E. A. Blaser, and A. M. Leslie, 2006, *Developmental Science*, *9*, p. 483. Copyright 2006 by Wiley-Blackwell. Reprinted with permission.

A third way in which the number of potentially relevant statistics could be reduced is by eliminating the "clutter" in visual scenes. There are two types of clutter: (a) the total number of combinations of *relevant* elements given a fixed number of elements, and (b) the total number of *statistics* that can be computed from the relevant elements. Typically, the number of relevant elements is much less than the number of elements, so if the learner could partition the scenes into relevant and irrelevant elements, it would substantially reduce the problem of too many statistics. The aforementioned Gestalt cues and hierarchy of stimulus salience are potentially effective in reducing the number of relevant elements. But how might a learner rapidly filter out the irrelevant statistics from the array of relevant elements?

We know that the second type of clutter—separating relevant statistics (signal) from irrelevant statistics (noise)—is difficult (Fiser & Aslin, in preparation). Even when the total number of elements is held constant, the more statistics that can be computed from combinations of these elements, the less effective are adults at extracting the coherent structures in the input set. The results from Younger and Cohen (1986) confirm this: when the correlations are defined across three features rather than across two features, even when the total number of features is held constant at five, infants are better able to learn the simpler statistics that define the conjunctions of features. One possible mechanism for limiting attention to the relevant statistics is suggested by a study of adults using speech streams similar to those of Saffran et al. (1996). Gebhart, Aslin, and Newport (2009) showed that when a particular set of statistics is learned initially, learning a second set of statistics from the

same inventory of elements is blocked. Thus, if infants are able to find the relevant statistics early in the learning process, perhaps by using Gestalt cues and intrinsic salience mechanisms, then the statistical clutter present in the larger set of images may be blocked from interfering with learning.

A fourth way in which the number of potentially relevant statistics could be reduced is by limitations in working memory. The rapid stimulus presentations used by Saffran et al. (1996: four syllables/s) and Fiser and Aslin (2002b: three-shape images every 2 s) place huge demands on the encoding of temporally ordered or spatially arranged elements. There is considerable empirical evidence that visual working memory is severely limited in infants. Under conditions of brief object occlusion, Kaldy and Leslie (2005) reported that working memory in 6-month-olds has a limit of one object. Ross-Sheehy, Oakes, and Luck (2003) reported that 6-month-olds can only keep track of rapid (500 ms) changes in a single object, whereas 10-month-olds succeed at set sizes of four objects (much like adults). And Oakes, Ross-Sheehy, and Luck (2006) reported that the ability to learn feature conjunctions develops rapidly between 6 and 7 months of age. Importantly, the capacity of working memory in adults is context specific (Alvarez & Cavanagh, 2004; see Zhang & Luck, 2008, for counter evidence). When the set of objects is easily encoded because they are highly discriminable, working memory is slightly greater than four, but when the set of objects is difficult to encode, working memory is substantially reduced to fewer than two.

How, then, might limitations in working memory work to the advantage of a statistical learner? Given a working memory capacity of two elements, infants who are briefly presented with more complex visual displays would be forced into a subsampling strategy—gathering a randomly selected portion of the entire stimulus scene. If the scenes were repeated a sufficient number of times to provide the same *aggregate* statistical information, then learning would occur but at a slower rate than if working memory had a greater capacity. Clearly, any loss of efficiency in learning due to subsampling is not a serious impediment when the overall structure to be learned is simple (as in artificial stimulus arrays). However, for a statistical learning mechanism to function effectively when confronted with the much higher structural complexity of natural language or natural images, working memory limitations must be overcome. Thus, without some other constraint, such as Gestalt cues or a hierarchy of stimulus salience, working memory limitations may not enable effective statistical learning except for structurally simple visual arrays. Baker, Olson, and Behrmann (2004) provide clear evidence in adults that the Gestalt cue of good continuation facilitates statistical learning.

A fifth way in which the number of potentially relevant statistics could be reduced is by the spatial scale of the stimuli, and the manner in which eye movements direct attention to the visual scene. Consider a fairly complex object like the "fribble" shown in Figure 6.5. When viewed at a far distance, such that the entire object falls within the central region of the retina, all of the relevant features can be accessed in a single fixation. In contrast, when viewed at a near distance, a series of eye movements is required to fixate the

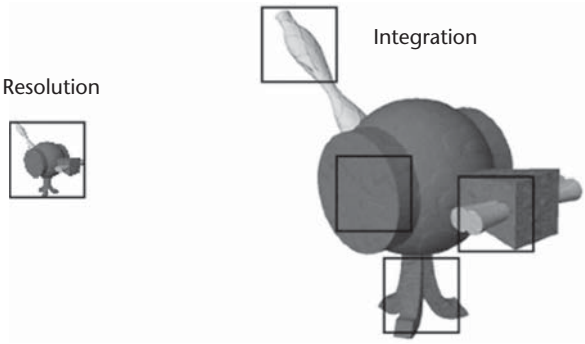

Figure 6.5. A sample image depicted at two spatial scales illustrating (a) how a small retinal image enables the entire object and all its component features to be accessed in a single fixation (represented by the black rectangle), and (b) how a large retinal image requires the component features of the object to be accesses by a series of fixations separated by eye movements. The smaller image requires excellent spatial resolution to perceive all the features, whereas the larger image requires integration of the more easily resolved features across eye movements. (See also figure in plate section.)
Source: Images reprinted from Michael J. Tarr, Brown University, http://www.tarrlab.org/.

many visual features. This example illustrates the tradeoff between spatial resolution (how many features are visible in the image) and feature integration (how many fixations are needed to access the features). If the image is too small, then many of the details of the features are inaccessible, but the features that are visible can be accessed in a single fixation. However, if the image is too large, then many more features are accessible but they can only be integrated by making a series of eye movements. Because eye movements take time to plan and execute, these delays place additional demands on working memory (see Najemnik & Geisler, 2005, for an optimal model of eye-movement control in a visual search task). Thus, if working memory is poor and only a very few features are resolvable within each fixation, the number of elements over which statistics can be computed is greatly reduced.

In unpublished work from my lab, using displays only slightly more complex than those used by Fiser and Aslin (2002b: see Figure 6.2), infants fail to extract the statistical coherence of pairs of shapes when the number of shapes is increased from three to four. Thus, limitations in working memory, and the requirement for element integration across eye movements, given the relatively large size of the array, may limit the ability to access even simple statistical coherence among shapes. Perhaps if additional exposure had been provided to the infants, a subsampling strategy would eventually have enabled effective statistical learning of the shape-pairs. However, all of the infants in this unpublished study had met the standard infant-controlled criterion for habituation, suggesting that this criterion is not necessarily a good measure of stimulus encoding when presented with many (16) rapid (2 s) images.

SUMMARY AND CONCLUSIONS

Younger and Cohen (1983) established a "statistical" approach to visual category formation by asking whether infants are sensitive to the correlations among features in 5-feature objects. Their results provided compelling evidence that feature correlations are not extracted efficiently until after seven months of age. Fiser and Aslin (2002b) elaborated on this statistical learning approach by characterizing *which* statistics are used to bind adjacent elements in multi-element scenes.

A key question confronting a statistical learning approach is how infants extract just the right statistics from the vast array of potential statistics available in even simple visual stimuli, rather than being overwhelmed by the task of computing both relevant and irrelevant statistics. Five constraints on statistical learning were summarized: Gestalt cues, a hierarchy of stimulus salience, a first-in bias that blocks subsequent statistical learning, limited working memory, and the spatial scale at which stimulus features are accessible during a single fixation. Future research should explore the effectiveness and developmental timing of these potential constraints on statistical learning to explain how infants succeed in grouping visual elements into spatially coherent features.

The statistical learning approach also has the potential to account for the extraction of nonadjacent feature correlations that serve to define categories. Category formation requires extracting features, recognizing the dimensions along which features vary, and filtering the relevant from the irrelevant dimensions. The general paradigm pioneered by Younger and Cohen (1983) has stood the test of time, and should lead to a productive reexamination of category formation using the quantitative designs of more recent statistical learning studies.

REFERENCES

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111.

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science*, *15*, 460–466.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 115–147.

Bulthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*, 247–260.

Civan, A., Teller, D. Y., & Palmer, J. (2005). Relations among spontaneous preferences, familiarized preferences, and novelty effects: Measurements with forced-choice techniques. *Infancy*, *7*, 111–142.

Cohen, L. B., Chaput, H. H., & Cashon, C. H. (2002). A constructivist model of infant categorization. *Cognitive Development*, *17*, 1323–1343.

Fiser, J., & Aslin R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.

Fiser, J., & Aslin R. N. (2002a). Statistical learning of higher order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458–467.

Fiser, J., & Aslin, R. N. (2002b). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*, 15822–15826.

Fiser, J., & Aslin, R. N. (in preparation). The effect of background statistics on visual structure learning.

Gebhart, A., Aslin, R. N., & Newport, E. L. (2009). Changing structures in mid-stream: Learning along the statistical garden path. *Cognitive Science*, *33*, 1087–1116.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192.

Gerhardstein, P., Kovacs, I., Ditre, J., & Feher, A. (2004). Detection of contour continuity and closure in three-month-olds. *Vision Research*, *44*, 2981–2988.

Kaldy, Z., Blaser, E. A., & Leslie, A. M. (2006). A new method for calibrating perceptual salience across dimensions in infants: The case of color vs. luminance. *Developmental Science*, *9*, 482–489.

Kaldy, Z., & Leslie, A. M. (2005). A memory span of one? Object identification in 6.5-month-old infants. *Cognition*, *57*, 153–177.

Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, *15*, 483–524.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*, B35–B42.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–391.

Oakes, L. M. Ross-Sheehy, S., & Luck, S. J. (2006). Rapid development of feature binding in visual short-term memory. *Psychological Science*, *17*, 781–787.

Quinn, P. C., & Bhatt, R. S. (2005). Learning perceptual organization in infancy. *Psychological Science*, *16*, 511–515.

Ross-Sheehy, S., Oakes, L. M., & Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Development*, *74*, 1807–1822.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Child Development*, *54*, 858–867.

Younger, B. A., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development*, *57*, 803–815.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–235.