Research Article

# UNSUPERVISED STATISTICAL LEARNING OF HIGHER-ORDER SPATIAL STRUCTURES FROM VISUAL SCENES

József Fiser and Richard N. Aslin

*Department of Brain and Cognitive Sciences and Center for Visual Science, University of Rochester*

**Abstract—***Three experiments investigated the ability of human observers to extract the joint and conditional probabilities of shape co-occurrences during passive viewing of complex visual scenes. Results indicated that statistical learning of shape conjunctions was both rapid and automatic, as subjects were not instructed to attend to any particular features of the displays. Moreover, in addition to single-shape frequency, subjects acquired in parallel several different higher-order aspects of the statistical structure of the displays, including absolute shape-position relations in an array, shape-pair arrangements independent of position, and conditional probabilities of shape co-occurrences. Unsupervised learning of these higher-order statistics provides support for Barlow's theory of visual recognition, which posits that detecting "suspicious coincidences" of elements during recognition is a necessary prerequisite for efficient learning of new visual features.*

Humans' visual experience consists of many complex spatiotemporal events that, under familiar circumstances, are rapidly and effortlessly interpreted with very few errors.[1] However, in an unfamiliar environment, such as a tropical rainforest or an exotic marketplace, complex scenes containing numerous three-dimensional objects are initially difficult to interpret. Yet, after a brief period of familiarization, observers are able to identify many exemplars of the initially unknown objects in these unfamiliar scenes. It has been suggested that this fundamental process of observational learning, which occurs automatically and without instruction, is what enables human infants to make sense of their visual environment during early development (Gibson, 1969). This learning process is also viewed as the general basis by which adults extract invariant visual features from multiple exemplars of a class of objects (Helmholtz, 1910/1925; Hochberg, 1981).

Experiments investigating the nature of this ability have all used feedback during training, and have found that observers are proficient at extracting components from exemplars of initially unknown scenes; they become "experts" in a given visual domain (e.g., Gauthier & Tarr, 1997; Quinn, Palmer, & Slater, 1999). Clearly, any learning during such a training process relies crucially on statistical components of the scenes, and the learned features can be characterized by statistics ranging from simple element frequency to higher-order spatial-temporal structures. Yet it has long been argued that when feedback is not provided during training, statistical learning alone cannot explain the ability to develop descriptors of visual experiences, or component features of objects, because of the prohibitive complexity of this unsupervised learning task, variously referred to as the "curse of dimensionality" (Duda & Hart, 1973) or the "combinatorial explosion" (von der Malsburg, 1995) problem. However, a clear empirical test of the adequacy or inadequacy of a statistical-learning mechanism, one that demonstrates which attributes are acquired in an unsupervised visual perception task, has not been reported previously.

We examined this issue by having observers view complex scenes to determine which statistics of visual features they naturally became sensitive to in the absence of feedback. The computational problem in all such statistical-learning tasks is to determine whether the simultaneous appearance of two features in a particular spatial relation, as indicated by their joint probability, $P(\mathrm{A}, \mathrm{B})$, is merely a random co-occurrence or a significant feature signaling an important underlying structure (Atick, 1992; Barlow, 1989, 1990). If the joint probability of a particular feature co-occurrence is low, then it would seem likely that those two features are unrelated. But over a large number of exemplars, nearly all joint probabilities will be low because of the large number of independent features in complex scenes. Thus, some robust decision criterion must be adopted so that relatively high joint probabilities among a sea of low joint probabilities are attended to and learned.

We investigated the ability of human observers to extract joint probabilities by creating a simple, highly structured set of exemplars in which all of the statistics among a discrete set of features (shapes) were under precise control. For the displays, we created a large number of visual scenes, each consisting of 6 two-dimensional shapes, selected from a set of 12, arranged on a fixed grid (Fig. 1). The shapes were chosen to be sufficiently complex that they were both easily discriminable and unfamiliar. Moreover, the complexity of the shapes prevented the automatic emergence of new shapes from the combination of adjacent low-level visual features (e.g., a line from a series of dots). Finally, the grid, which was continuously present during the experiment, was used to minimize uncertainty about the absolute and relative spatial positions of the shapes. Thus, in our task, observers could learn the higher-order spatial relations among the shapes only by computing one or more statistics from the distribution of the exemplars and not from Gestalt laws of organization or from prior familiarity with the shapes.

The paradigm we employed is superficially similar to two other lines of research that bear on related issues. Johnson, Peterson, Yap, and Rose (1989) presented a total of seven letters and digits simultaneously within a 4 × 4 grid. After viewing 128 grids (scenes) in which element frequency ranged from 2 to 24 instances, subjects showed sensitivity to element frequency, even when instructed only to look at the elements in each scene. Thus, a low-level statistic (relative frequency or probability) can be extracted from multielement arrays during passive viewing. Chun and Jiang (1998, 1999) examined the influence of context—the background elements that surround a single target element—in a visual search task. Subjects showed faster search times to targets embedded in familiar contexts than to targets embedded in novel contexts. Contextual familiarity emerged from the repeated pairing of targets in specific background arrays, even though

Address correspondence to Richard N. Aslin, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627; e-mail: aslin@cvs.rochester.edu.

[1] We are not concerned here with phenomena that are due to rapid temporal changes of scenes under specialized circumstances, such as change blindness and attentional blink.

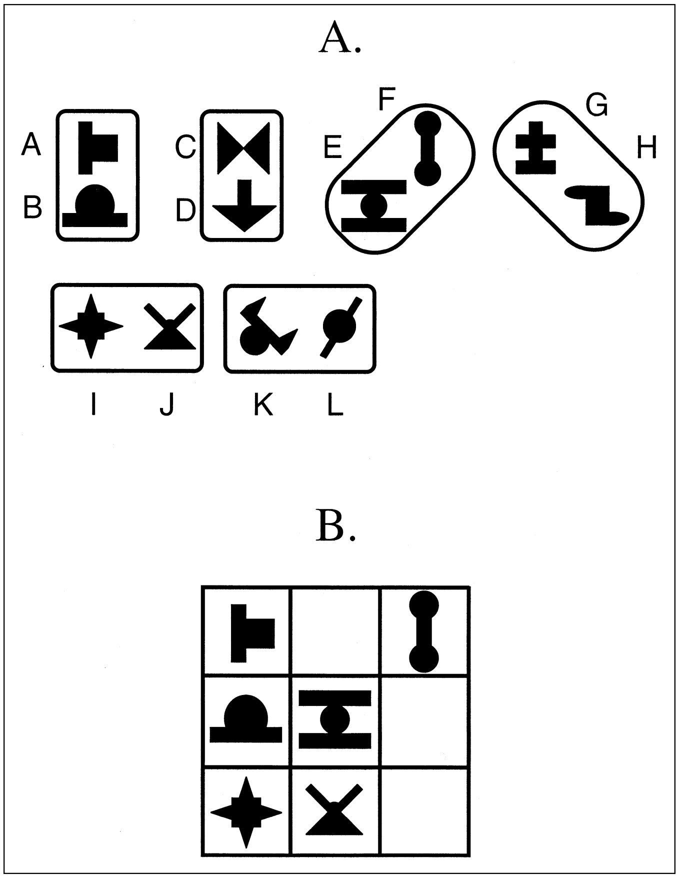

**Fig. 1.** The 12 basic shapes used in all of the experiments (a) and a typical scene used in Experiments 1 and 2 (b). As illustrated in (a), the shapes were arranged into six base pairs, with each base pair defined by two shapes and a particular spatial relation between them. (The outlines around the pairs are shown here for illustrative purposes only.) Assignment of shapes to pairs was randomized across subjects, but for each subject (as shown here), two base pairs were organized into each of three orientation groupings (horizontal, vertical, oblique). Each of the scenes was composed from three base pairs.

subjects were instructed to seek a particular target element and were unaware of the consistent mapping between targets and contexts. These results demonstrate that a specific shape, which is the explicit object of attention, can be associated implicitly with a multielement background array. However, it remains unclear whether higher-order statistics can be learned from multielement arrays when there is no explicit object of attention, and which particular statistics are used during this process of observational learning.

## EXPERIMENTS 1 AND 2

### Method

*Subjects*

Separate groups of 20 naive subjects participated in the two experiments. The subjects were undergraduates at the University of Rochester who were paid $6.00 for their participation.

*Stimuli*

Twelve arbitrary complex shapes were created in the Canvas drawing program from simple two-dimensional figures. The shapes were black on a white background and were displayed within a 3 × 3 grid. The maximum height and width of the shapes were scaled to be equal, and to be half of the extent of each cell in the grid. The stimuli were presented on a 21-in. Sony Trinitron 500PS monitor at 1024 × 728 resolution from a 1-m viewing distance. The extent of the 3 × 3 grid was 13.7°, and the maximum size of the shapes was 2.29°. Stimuli were presented on a Macintosh G3 computer using Matlab and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

*Design*

Each experiment consisted of two phases: familiarization and test. During the familiarization phase of each experiment, but unknown to the subjects, the 12 shapes were organized into six base pairs, each base pair consisting of two given shapes in a particular spatial relation (Fig. 1a). Base pairs can be thought of as objects or rigid parts, in that if one of the elements of a base pair appeared in a given scene during familiarization, the other element always appeared in an invariant spatial relation to it. The specific assignment of the 12 shapes to the six base pairs was randomized across subjects to ensure that specific shape pairs were not unusual and more (or less) easily learned.

As shown in Figure 1a, the six base pairs were organized into three orientation groupings: horizontal, vertical, and oblique. The scenes were created by selecting one base pair from each of the three orientations and then randomly positioning these three base pairs in the 3 × 3 grid so that each base pair would neighbor at least one of the other pairs (Fig. 1b). This use of spatial adjacency ensured that the learning of base pairs was not facilitated by obvious segmentation cues. In Experiments 1 and 2, the foregoing constraints on how base pairs could be arranged in the 3 × 3 grid limited each base pair to four possible locations within the grid, and created a total of 144 possible scenes, which were presented to each subject, in random order, during familiarization. Because each base pair appeared in half of the scenes, and the two shape elements of a base pair always appeared together, the probability of appearance for each element, as well as the joint probability of the two shapes in each of the six base pairs, was .50. The configuration of the base pairs resulted in accidental co-occurrences when one shape of one base pair was located next to another shape of a different base pair. The joint probability of such coincidental non-base pairs was fairly homogeneous and typically less than .02, a value much smaller than that of the base pairs.

*Procedure*

During familiarization, subjects saw each of the 144 possible scenes only once (a total of 7 min). Each scene was shown for 2 s, with a 1-s pause between scenes. This scene duration is consistent with the report (Johnson et al., 1989) that judgments about the relative frequency of seven elements in a 4 × 4 grid minimally require a 2-s inspection time. Subjects were told to pay attention to the continuous sequence of scenes so that they would be able to answer some simple questions after the familiarization phase. No further instructions were given, thereby ensuring that the subjects were unaware of the aspect of the changing scenes to which they should attend. There was a 3-min break between the familiarization and the test phase.

After the familiarization phase, a temporal two-alternative forced-choice (2AFC) test phase was conducted. During this test phase, a base pair and a non-base pair were shown sequentially in particular positions in the 3 × 3 grid. Each pair was presented for 2 s, with a 1-s pause between pairs. Subjects had to press a computer key ("1" or "2") depending on which of the two test patterns they judged to be more familiar. Test trials were individually randomized for each subject, and base pairs and non-base pairs were counterbalanced within a test session.

In Experiment 1, the non-base pairs were constructed so that neither of the individual shapes of each pair had appeared in the tested grid position during the familiarization phase. Figure 2a shows an example of a non-base pair in the sample test scene for Experiment 1. This pair is composed of shapes I and C (see Fig. 1a). However, during familiarization, shape I could never appear in the right-most column of the grid, and shape C could never appear in the lowest row of the grid. Thus, this non-base-pair test scene (I above C) could not appear during familiarization. Therefore, subjects could rely on shape position as well as shape-shape relational information in their 2AFC comparison of base pairs with non-base pairs.

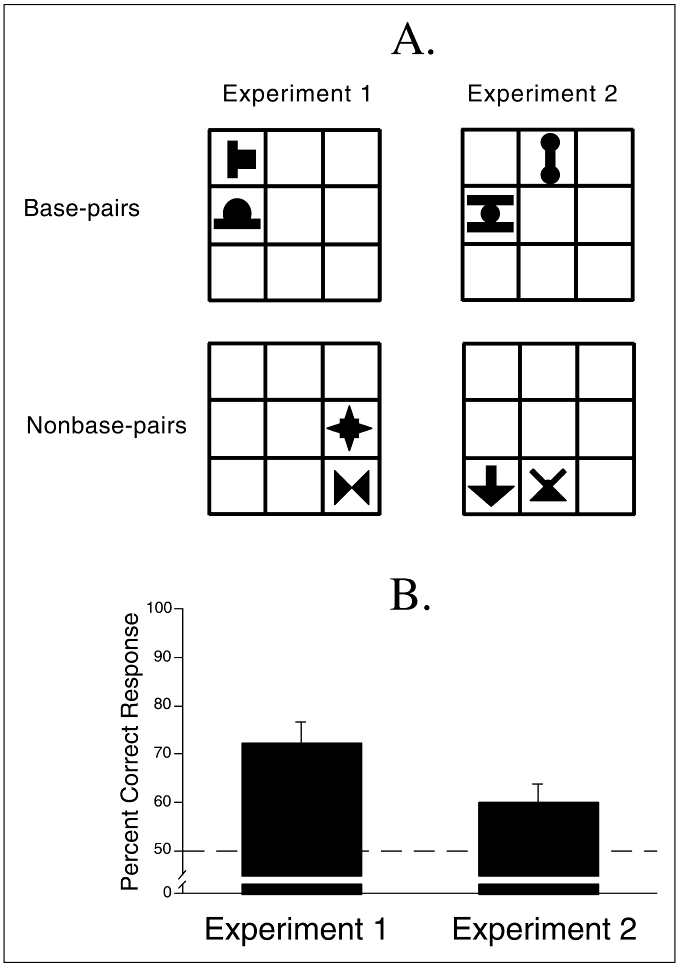

**Fig. 2.** Sample test trials (a) and results (b) from Experiments 1 and 2. The sample trials show base pairs (top row) and non-base pairs (bottom row), in particular grid positions. In the graph giving the results, the *y*-axis is truncated below 50%, which was chance performance in both experiments. Error bars indicate standard errors of the mean.

In Experiment 2, the individual shapes of each non-base pair had appeared in the tested positions of the grid during familiarization as frequently as the individual shapes of the base pair had appeared in the tested positions, but the two shapes of the non-base pair had never occurred together in the tested spatial arrangement in the scenes presented during the familiarization phase (Fig. 2a). For example, during familiarization, shape D of Figure 1a could appear in the lower left corner of the grid, and shape J could appear in the second column of the lowest row of the grid, but the non-base pair (D left of J) shown in Figure 2a could not appear during familiarization because when shape J appeared, shape I was always to its left. Thus, subjects could not rely on the positional information of individual shapes in the grid to choose the base pairs in the 2AFC test of Experiment 2.

### Results and Discussion

The results of the first two experiments are shown in Figure 2b. In Experiment 1, subjects reliably distinguished between base pairs and non-base pairs when the individual shapes in the non-base pairs were presented in grid locations inconsistent with the familiarization phase, $t(19) = 6.16, p < .0001$. Subjects selected the base pair as the familiar pattern significantly more often than the non-base pair despite the fact that the individual shapes, whether contained in the base pair or in the non-base pair, appeared an equal number of times in the familiarization scenes [i.e., $P(A) = P(B) = \ldots P(L)$]. Thus, subjects learned, without feedback, higher-order statistical characteristics of the scenes that went beyond the frequency of individual shapes.

There are two types of higher-order statistics that the subjects could have used to distinguish base pairs in Experiment 1. First, they could have learned that certain individual shapes never appeared in certain grid positions (i.e., the joint probability of shape and grid position). Second, they could have learned that the joint probability of shape-pair co-occurrence was much higher for base pairs than for non-base pairs, regardless of their position in the grid. Experiment 2 was designed to test whether subjects are able to perform the task even when the first source of information is not available. As shown by the second bar in Figure 2b, subjects judged the base pairs as more familiar than the non-base pairs even when they could not rely on absolute grid position to distinguish between base pairs and non-base pairs, $t(19) = 3.28, p < .005$. Because of the identical training and instructions in the two experiments, and because participants did not know which type of test items would be presented, the results of the two experiments demonstrate that participants learned both position-dependent and position-independent higher-order statistics of the training scenes, in parallel and in 7 min or less. However, subjects' performance was significantly weaker in Experiment 2 than in Experiment 1, $t(38) = 2.32, p < .03$, suggesting that the test is easier when the absolute position of the shapes, as well as their joint probability of co-occurrence, can be utilized.

## EXPERIMENT 3

The problem with exclusive reliance on joint probabilities is that they often do not signal the predictiveness of feature co-occurrences. Consider, for example, a simple case in which there are five features: $\alpha$, $\beta$, X, Y, and Z. Features $\alpha$ and $\beta$ always co-occur (they never occur in isolation), but their joint probability is relatively low. In contrast, features X and Y both co-occur with feature Z, but because of variations in the frequency of co-occurrence, $P(\mathrm{X}, \mathrm{Z})$, $P(\mathrm{Y}, \mathrm{Z})$, and $P(\alpha, \beta)$ are equal. If joint probability were the sole criterion for attending to and learning about the reliability of co-occurrences, then the more predictive relation between α and β would be acquired no more readily than the less predictive relations between X and Z and between Y and Z. A statistic that better captures this predictive relation is conditional probability, $P(\beta|\alpha)$, because it normalizes the joint probability of the two features, $P(\alpha, \beta)$, with respect to the probability of the predicting feature, $P(\alpha)$. In this simple example, all joint probabilities are equal, but the conditional probability of β given α is 1.00, whereas the conditional probabilities of Z given X and Z given Y are .50. Thus, conditional probabilities provide a superior metric for extracting invariant properties from highly variable visual scenes (Atick, 1992; Barlow, 1989, 1990).

In Experiment 3, the joint probabilities of some of the base pairs and some of the non-base pairs were equated by varying the relative frequency of the different base pairs during the familiarization phase. Specifically, the frequency of two of the six base pairs was doubled, and extra constraints were added so that when these more frequent base pairs appeared together in the scene, their relative position would be the same in half of those scenes. As a result, the number of "accidental" co-occurrences of shapes across the two more frequent base pairs (cross-pairs) was equal to the number of co-occurrences of the two shapes within the rare base pairs. In other words, the joint probabilities of the cross-pairs and some of the base pairs were identical; these test items made up the *frequency-balanced* condition. Thus, in contrast to Experiments 1 and 2, the differences in joint probabilities between base pairs and cross-pairs could not be utilized during the test phase of this experiment. Nevertheless, the conditional probability of the rare base pairs was 1.0 (when one of the shapes appeared, the other shape always appeared in the proper relative grid position), whereas the conditional probability of cross-pairs was .50, because only half of the time during the familiarization phase did the second shape appear in conjunction with the first shape (Fig. 3a).

### Method

In Experiment 3, an independent group of 20 subjects viewed a 25-min sequence of scenes,[2] with the same instructions as in Experiments 1 and 2. Each scene was composed of three base pairs, as in Experiments 1 and 2, but the size of the grid was increased to 5 × 5 to allow for the added constraints on base-pair location during familiarization (see Fig. 3a). A total of 212 unique scenes was presented twice (in random order), with a scene duration of 2 s. The extent of the 5 × 5 grid was 11.4°, and maximum size of each shape was 1.14°. Some of the base pairs (depicted on the left in Fig. 3a) had a higher frequency of appearance than other base pairs. This resulted in equal frequency of appearance of the two cross-pairs (depicted on the right in Fig. 3a) and the two rare base pairs (not shown). However, the conditional probability of these frequency-balanced cross-pairs was only half that of the rare base pairs.

After the familiarization phase, subjects completed two temporal 2AFC tests, one with shape pairs and one with single shapes. In the pair-based 2AFC test phase, subjects indicated which of two shape pairs, one rare base pair and one frequency-balanced cross-pair, was more familiar. In the single-shape 2AFC test, subjects judged which shape was more frequent during the familiarization phase. In contrast to the test displays in Experiments 1 and 2, all single-shape and shape-pair test displays in Experiment 3 were located in the center of the grid to eliminate absolute grid position as a relevant source of information.

### Results and Discussion

Because the co-occurrence frequency of base pairs and cross-pairs was equated in Experiment 3, the only information available for the subjects to distinguish between these two test items was the predictability between individual shapes, that is, the conditional probability between the elements of the shape pairs. As shown in Figure 3b, subjects were sensitive to this information, as indicated by their reliable selection of the base pairs as more familiar than the frequency-balanced cross-pairs, $t(19) = 3.53$, $p < .005$. In addition, the results of the single-shape task demonstrated that while subjects were extracting higher-order statistics, they also maintained sensitivity to the frequency of the individual shapes, $t(19) = 6.61$, $p < .0001$.

## GENERAL DISCUSSION

In three experiments, we found that subjects spontaneously and in parallel learned first-order and a variety of higher-order statistics from the spatial arrangement of shapes in scenes. The learned statistics ranged from single-shape frequency, to absolute shape position in an array, to shape-pair arrangement independent of position, and finally to position-independent conditional probability of shape co-occurrence. The learning was unsupervised because no instructions directed the subjects' attention to the underlying base-pair structure or any other characteristic of the scenes. In addition, most subjects spontaneously reported that they were unaware of the relation between the test trials and the familiarization displays, and felt that they were often guessing.

These results are important because virtually every recent model aimed at describing how visual recognition operates relies on the learning of higher-order statistical features from complex scenes (Barlow, Kaushal, & Mitchison, 1989; Dayan, Hinton, Neil, & Zemel, 1995; Mel, 1997; Penev & Atick, 1996; Poggio & Edelman, 1990; Riesenhuber & Poggio, 1999), yet there is only indirect evidence that humans are able to extract the statistics required for carrying out such learning efficiently, unless they are given some form of feedback to constrain the statistical-learning process. Especially instructive are the frequency-balanced shape-pair results from Experiment 3 because they demonstrate statistical learning that is more complex than simple sensitivity to the frequency of appearance of single shapes or pairs of shapes.[3] A number of studies have reported that humans are sensitive

[2] Pilot testing revealed that shorter durations of familiarization were insufficient for above-chance learning of conditional probabilities for frequency-balanced shape pairs.

[3] Our goal was to provide subjects with a variety of precisely controlled statistics in static, presegmented, spatially adjacent, multishape displays. Undoubtedly, objects in the real world contain a variety of statistics that are less well controlled. However, it is not clear whether this greater real-world variability is advantageous (assisted by correlated cues) or disadvantageous (harmed by unsegmented features). In the absence of detailed information about the distribution of image features in natural scenes, our results are nevertheless important in demonstrating that sophisticated statistical-learning mechanisms are available and capable of operating in real time to rapidly extract shape co-occurrences.

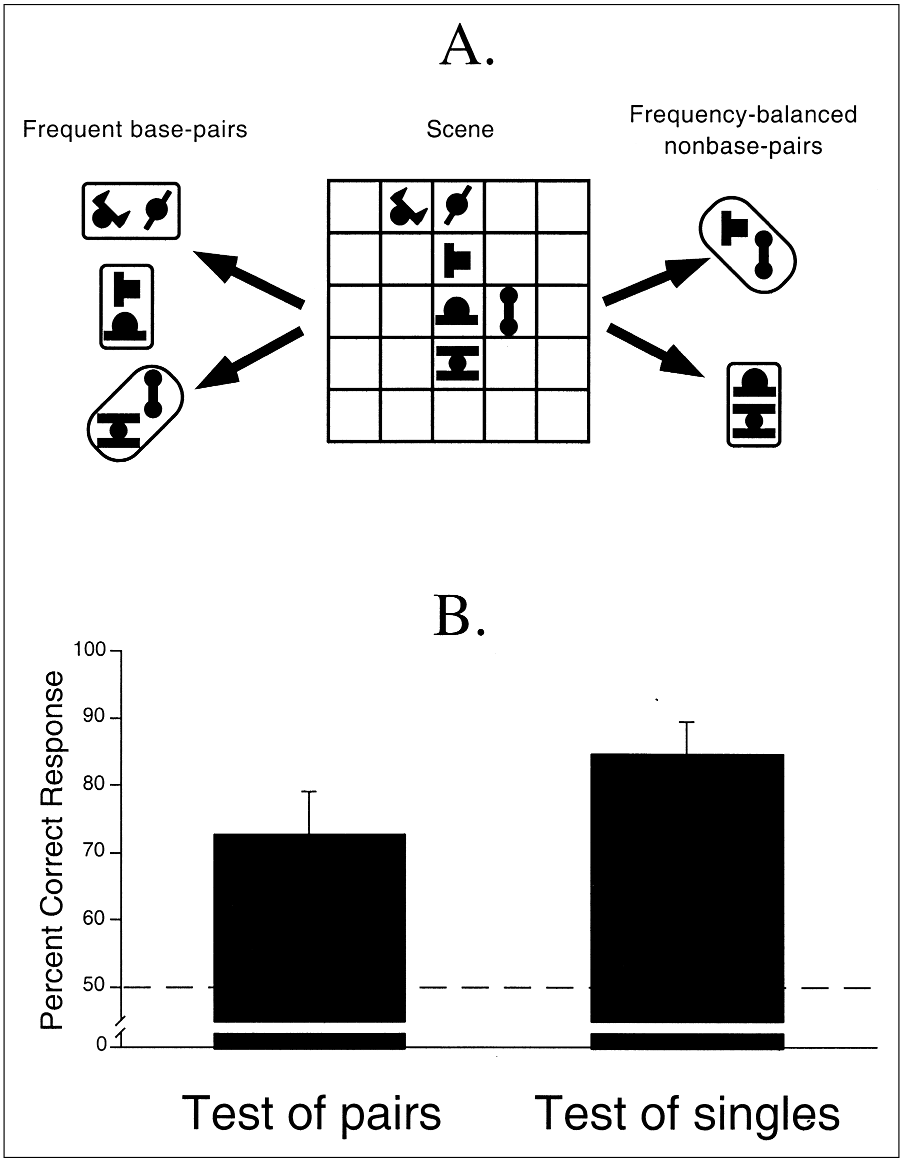

**Fig. 3.** A sample scene (a) and results (b) from Experiment 3. The sample scene illustrates frequent base pairs (on the left) and frequency-balanced non-base pairs (on the right). The graph shows the results of the pair and the single-element test comparisons. In the graph giving the results, the *y*-axis is truncated below 50%, which was chance performance in both tests. Error bars indicate standard errors of the mean.

to appearance frequencies (e.g., Hasher & Zacks, 1984; Johnson et al., 1989), but unsupervised learning of conditional probabilities has not been reported before in the visual domain, and even in the auditory domain it has been reported only in the context of sensitivity to transitional probabilities of phonemes in infants (Aslin, Saffran, & Newport, 1998).[4]

The present results support the conjecture that the brain performs effective associative learning by spontaneously and automatically extracting independent components from complex visual scenes (Barlow, 1990; Helmholtz, 1910/1925; Hinton & Ghahramani, 1997). This possibility is in agreement with the suggestions and physiological findings that the primary visual cortex (V1) generates a sparse, efficient representation of visual scenes (Field, 1994; Vinje & Gallant, 2000), whereas areas beyond V1 compute more complex, higher-order descriptors based on this early representation (Logothetis & Sheinberg, 1996; Tanaka, 1996). Although we have no direct evidence about the neural mechanisms that mediate statistical learning in the multishape task reported here, it is likely that the learning in our task occurs

[4] Other studies from our research group (Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Saffran, Newport, & Aslin, 1996) did not employ designs that separated the effects of conditional probabilities from the frequency of n-grams.

at a higher level in the visual system than the low-level (e.g., Fahle & Edelman, 1993; Karni & Sagi, 1993) and mid-level (Ahissar & Hochstein, 1997; Liu & Weinshall, 2000) types of perceptual learning reported in the literature.

Our results, as well as those of Chun and Jiang (1998, 1999), demonstrate that higher-order statistics among spatial arrays of shapes can be extracted involuntarily, because our subjects were not instructed to attend to particular sources of information. However, our results are not necessarily in conflict with other studies reporting that some aspects of attention or feedback may be needed for learning during visual tasks (Ahissar & Hochstein, 1993; Johnson et al., 1989; Shiu & Pashler, 1992). For example, our subjects were asked to fixate and attend to the visual displays, and we suspect that failure to do so consistently would have resulted in a significant decline in discrimination performance. However, we do suggest that once subjects engage their general attention to scenes, statistical-learning processes are automatically activated, and these processes are sufficient to extract a variety of lower- and higher-order statistics. We further suggest that inefficient deployment of attention, or explicit generation of incorrect hypotheses about the underlying structures, may interfere with this automatic statistical-learning process. Clearly, statistical learning is not the only mechanism available to learners, even under implicit conditions, but we believe that statistical-learning mechanisms are widely available across modalities, domains, and species (Hauser, Newport, & Aslin, 2001).

If, as suggested by our results, subjects rely on statistical learning in processing unknown scenes, how can one reconcile this with the combinatorial-explosion problem, which states that there are not enough exemplars to learn the necessary higher-order probabilities in complex natural scenes? Recent large-scale computer simulations have shown that the most efficient object recognition system utilizes features at many different levels of complexity, ranging from very simple features to features constructed by conjunctions based on higher-order conditional probabilities (Mel & Fiser, 2000). However, if the more complex features are recruited only as needed—that is, only when lower-order features fail to provide sufficient information to solve a particular task—then the minimally sufficient number of features decreases exponentially as their complexity increases. Therefore, a full-blown search through high-dimensional feature spaces is not necessary to solve the recognition problem. These results imply that if humans are able to access increasingly higher-order statistics from scenes, but use only a constrained set of these statistics to extract higher-order features, learning may avoid the combinatorial-explosion problem (Geman, Bienenstock, & Doursat, 1992). Our findings, that subjects learned both position-dependent and position-independent statistics in the first two experiments and learned the frequency of individual elements as well as conditional probabilities for shape pairs in the third experiment, support the hypothesis that the human brain indeed employs such a strategy to reduce the combinatorial-explosion problem.

**Acknowledgments—**This research was supported by a postdoctoral fellowship to J.F. from the James S. McDonnell Foundation (96-32) and by a research grant to R.N.A. from the National Science Foundation (SBR-9873477). We are grateful to Elizabeth Gramzow, Rebecca DaMore, Conni Augustine, Rachel Heafitz, and Koleen McCrink for testing the subjects, and to Ruskin Hunt, Robert Jacobs, David Knill, William Merigan, Alex Pouget, Stéphane Rainville, and James Ringo for critical comments on an earlier draft.

## REFERENCES

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences, USA*, *90*, 5718–5722.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*, 401–406.

Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Atick, J.J. (1992). Could information theory provide an ecological theory for sensory processing? *Network: Computation in Neural Systems*, *3*, 213–251.

Barlow, H. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, *30*, 1561–1571.

Barlow, H.B. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.

Barlow, H.B., Kaushal, T., & Mitchison, G. (1989). Finding minimum entropy codes. *Neural Computation*, *1*, 412–423.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 443–446.

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71.

Chun, M.M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*, 360–365.

Dayan, P., Hinton, G.E., Neil, R.M., & Zemel, R.S. (1995). The Helmholtz machine. *Neural Computation*, *7*, 889–904.

Duda, R.O., & Hart, P.E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Fahle, M., & Edelman, S. (1993). Long-term learning in Vernier acuity: Effects of stimulus orientation, range and of feedback. *Vision Research*, *33*, 397–412.

Field, D. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Gauthier, I., & Tarr, M.J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*, 1673–1682.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Gibson, E.J. (1969). *Principles of perceptual learning and development.* New York: Appleton-Century-Crofts.

Hasher, L., & Zacks, R.T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *39*, 1372–1388.

Hauser, M.D., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*, B53–B64.

Helmholtz, H., von. (1925). *Treatise on physiological optics* (J.P.C. Southall, Ed.). Washington, DC: Optical Society of America. (Original work published 1910)

Hinton, G.E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London B*, *352*, 1177–1190.

Hochberg, J. (1981). On cognition in perception: Perceptual uncoupling and unconscious inference. *Cognition*, *10*, 127–134.

Johnson, M.K., Peterson, M.A., Yap, E.C., & Rose, P.M. (1989). Frequency judgments: The problem of defining a perceptual event. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 126–136.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, *365*, 250–252.

Liu, Z., & Weinshall, D. (2000). Mechanisms of generalization in perceptual learning. *Vision Research*, *40*, 97–109.

Logothetis, N.K., & Sheinberg, D.L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Mel, B.W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, *9*, 777–804.

Mel, B.W., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation*, *12*, 731–762.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Penev, P.S., & Atick, J.J. (1996). Local feature analysis: A statistical theory for object representation. *Network: Computation in Neural Systems*, *7*, 477–500.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*, 263–266.

Quinn, P., Palmer, V., & Slater, A.M. (1999). Identification of gender in domestic-cat faces with and without training: Perceptual learning of a natural categorization task. *Perception*, *28*, 749–763.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition. *Nature Neuroscience*, *2*, 1019–1025.

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

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Shiu, L.P., & Pashler, H. (1992). Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Perception & Psychophysics*, *52*, 582–588.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139.

Vinje, W.E., & Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*, 1273–1276.

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, *5*, 520–526.

(RECEIVED 8/12/00; REVISION ACCEPTED 2/15/01)