Journal of Experimental Psychology: General
2005, Vol. 134, No. 4, 521–537

0096-3445/05/$12.00 DOI: 10.1037/0096-3445.134.4.521

Encoding Multielement Scenes: Statistical Learning of Visual Feature Hierarchies

József Fiser and Richard N. Aslin
University of Rochester

The authors investigated how human adults encode and remember parts of multielement scenes composed of recursively embedded visual shape combinations. The authors found that shape combinations that are parts of larger configurations are less well remembered than shape combinations of the same kind that are not embedded. Combined with basic mechanisms of statistical learning, this embeddedness constraint enables the development of complex new features for acquiring internal representations efficiently without being computationally intractable. The resulting representations also encode parts and wholes by chunking the visual input into components according to the statistical coherence of their constituents. These results suggest that a bootstrapping approach of constrained statistical learning offers a unified framework for investigating the formation of different internal representations in pattern and scene perception.

Keywords: perceptual learning, implicit memory, scene perception, visual features, chunking

The purpose of the current work is to investigate how human adults develop new visual representations to encode and recognize both familiar and novel objects and scenes in situations in which the visual features in the input are hierarchically structured. We began by utilizing a statistical learning framework, in which two-dimensional information was encoded by the visual system using a modest number of memory traces, which in turn were combined in a variety of ways to form higher order visual representations. The focus of our experiments was on the origin of these representations, their relation to the statistical regularities of the visual environment, and the feasibility of a statistical learning mechanism for acquiring such representations from a hierarchically organized input so that they were sufficient for recognition. In a series of five experiments, we investigated how humans encode and remember higher order visual shape combinations by simple observation of multielement scenes without performing any particular task or receiving specific feedback, a situation that mimics the process of visual encoding in a natural context.

Our program of research built on the dominant approach to developing visual object representations, which proposes a transformation of early neural codes into more complex specialized representations, or "features," that retain only part of the input information as it moves to higher levels of representation in the visual cortex (Biederman, 1987; Marr, 1982). This required the selection of an appropriate set of features from the large pool of potential candidates by using powerful learning mechanisms that can, over a range of time scales (from seconds to years), acquire information about objects, surfaces, and scenes at multiple spatial scales and develop feature detectors of different complexities to solve particular tasks.

This approach raised two fundamental questions. First, what determines the elementary feature detectors used by the visual system to represent the whole or the different parts and components of a given object or scene? Second, what is the computationally tractable process that enables new feature detectors to develop so that visual representations remain efficient without losing the capacity to capture unfamiliar inputs?

Regarding the first question, it is clear that the visual system can identify and, therefore, must be able to represent and remember a vast number of highly complex two-dimensional shapes, three-dimensional objects, and multiobject scenes. However, there are very few hypotheses about what sort of features might serve perception beyond the early stages of the visual pathway. Thus, it is not surprising that even less is known about whether and how detectors of these features develop based on visual experience. It is not even clear whether there is a hierarchy of visual features, and, if so, at what level of the visual feature hierarchy learning can take place. Moreover, despite demonstrations of perceptual learning in many visual tasks (Ahissar & Hochstein, 1997; Ball & Sekuler, 1982; Fiorentini & Berardi, 1980; Karni & Sagi, 1991; Poggio, Fahle, & Edelman, 1992; Ramachandran & Braddick, 1973), there is no consensus about what aspects of the visual input can or cannot be learned or whether perceptual learning is related to the normal process of developing higher order visual represen-

József Fiser and Richard N. Aslin, Department of Brain and Cognitive Sciences, University of Rochester.

This research was supported by Center for Visual Science Training Grant EY-07125, National Institute of Child Health and Human Development Research Grant HD-37086, National Science Foundation Grant SBR98-73477, and Center for Visual Science Core Grant EY-01319. We are indebted to Koleen Gochal-McCrink and Julie Markant as well as the students in the Infant Perception Lab for their assistance in the data collection process. Daeyeol Lee and Dave Knill provided helpful critical comments on a draft of this article.

Correspondence concerning this article should be addressed to József Fiser, who is now at the Department of Psychology and the Volen Center for Complex Systems, Volen Room 208, MS 013, Brandeis University, Waltham, MA 02454. E-mail: fiser@brandeis.edu

tations. Regarding the second question, any identifiable visual feature can be broken down into smaller subfeatures while remaining part of some larger superfeature. This high level of embeddedness means that the learning problem is defined in a very large dimensional space that, in turn, poses a hard constraint on the feasibility of the candidate learning mechanisms.

We address these two questions within a statistical learning framework, in which the development of a new feature is based on the gradual collection of evidence from a corpus of visual scenes. Therefore, in our view, developing a new feature detector and remembering a fragment of a previously seen, but unfamiliar, scene are intimately linked on a representational level. In this framework, a feature is defined as a subset of preexisting stored traces of visual signatures linked by past associations that the individual is able to remember when it is presented independently of its original context. A visual signature can be any previously acquired feature as well as simple measures of visual information in the scene, such as contrast discontinuities, light patches, color combinations, edge structures, and so forth. Thus, a feature can be the stored trace of an entire scene as well as that of the smallest individual signature of the scene or any combination of signatures of arbitrary complexity. We explore the general hypothesis that there is a limited set of general constraints on encoding that enables human adults to extract higher order visual features in a statistical manner, and we try to identify one of these general constraints that make statistical visual feature learning computationally plausible. We begin by highlighting a general computational aspect of statistical learning that shows why a statistical learning approach to extracting visual features requires constraints to render it plausible. We then introduce an experimental paradigm that enables us, independently of specific visual attributes, to ask fundamental questions about how remembering visual features is related to the statistical information in the input. Finally, we present the results from five statistical learning experiments that suggest a particular implicit strategy used by humans to derive visual memories of fragments from hierarchically organized multielement scenes. The results of these experiments suggest that a powerful constraint, embeddedness, is automatically activated when encoding complex visual inputs and enables a statistical learning mechanism to account for the chunking of the input into parts.

Computational Constraints on Statistical Learning

Visual images reside in a very high dimensional data space, but in an entire lifetime humans only encounter a very small and very specific subset of the images from this high-dimensional space (Field, 1994). For example, the set of images composed of only 10×10 pixels, each of which can have either a black or white luminance value, consists of 2^{100} unique images, which is about 20 orders of magnitude greater than the number of images that the visual system encounters in a lifetime.[1] For comparison, in any moment each retina forwards about 1 million (10^6) pixels of information to the brain. Thus, the problem of visual recognition is defined in an extremely large dimensional data space that defies straightforward learning methods, unless there are biases that guide the learning process. In addition to this problem of dimensional complexity, visual recognition does not operate on a set of static images but on a dynamic sequence of images that typically change several times per second (either because the stimulus itself changes or because the eyes move). Thus, visual perception is a serial process in that there is a nearly continuous updating of scenes containing many objects distributed over time.

The foregoing aspects of visual structure and visual processing impose strong constraints on how any statistical learning mechanism could be used in the domain of visual learning. Even if some of the correlations of subelements in a scene could be computed in parallel when searching for a feature, the amount of input data—in the case of vision the number of visual scenes and, within scenes, the number of co-occurring individual elements—required to decide what a feature might be grows exponentially with the number of added elements. In a large dimensional space, this leads to intractable computational demands.

This well-known problem of insufficient data for large learning problems is called the "curse of dimensionality" (Bellman, 1961), and it sets a hard limit on what can be learned in high dimensional spaces with a brute force approach. The only way to handle the curse of dimensionality is to reduce the size of the problem by decreasing the number of relevant dimensions within which new associations have to be learned (Geman, Bienenstock, & Doursat, 1992). The size of a learning problem can be reduced by imposing particular constraints on the learning process based on some assumptions about what should be learned. Because visual recognition works in a high dimensional space, and if humans use statistical learning to acquire new higher order features, then it can only operate successfully by imposing some general constraints on the learning process by the visual encoding and recognition system.

An Observational Learning Paradigm

We propose a three-component framework, using an observational learning paradigm, for investigating visual statistical learning in humans. The first component is that feature learning is an unsupervised, natural process that is quite different from the perceptual learning paradigm (with feedback) customarily used for studying human visual learning. The second is that the fundamental problem in learning higher order features is not that Gaussian noise obscures some of the elements, thereby preventing the correct underlying associations from being extracted. Rather, there are far too many possible combinations of elements that all could be potentially relevant higher order features, thereby creating a computational problem in the search for the features that are meaningful. The third is that to investigate the general mechanisms of visual learning, it is necessary to use stimuli with characteristics that cannot skew the results because of some low-level built-in constraint. Simple visual stimuli can form higher level features by recourse to already existing lower level grouping principles that have little to do with learning. We elaborate on each of these three issues in some detail.

[1] This is a rough approximation, assuming that humans can observe three new independent images per second because of saccadic eye movements, considering their life span is about 80 years, and not counting any idle time due to sleeping.

Visual learning is often studied in perceptual learning paradigms (Ahissar & Hochstein, 1997; Ball & Sekuler, 1982; Fahle & Poggio, 2002; Fiorentini & Berardi, 1980; Karni & Sagi, 1991; Poggio et al., 1992; Ramachandran & Braddick, 1973). In these paradigms, the individual is presented with a well-defined task that is explained verbally by the experimenter. With repetitive training (including feedback), the individual's performance improves. Although this paradigm is useful in characterizing the conditions necessary for skill learning, it is quite unrealistic as a model of acquiring higher order visual features in the natural environment: For the most part, higher order features are learned without a teacher or a task.

In our observational learning paradigm, the only task facing the individual is to pay attention to the scenes. The statistical complexity of the scenes is under tight experimental control and allows for higher order features to emerge. Because the goal of our studies is to investigate general aspects of statistical learning, we avoid stimuli that invoke specific mechanisms already implemented in the visual system that might mistakenly be interpreted as evidence of learning. Some of these special mechanisms are well known, such as greater sensitivity to horizontal than to oblique structures or the tendency to regularize contour shapes or fill in gaps. These special mechanisms are important for a full implementation of the many constraints on visual perception, and they should be investigated in their own right (Field, Hayes, & Hess, 2000; Kourtzi, Tolias, Altmann, Augath, & Logothetis, 2003). However, as an initial strategy for studying the most general constraints on visual learning, one needs a stimulus set in which the processing of the statistical relationships between subelements can be investigated independent of these other special-purpose mechanisms.

Paradoxically, this goal can best be achieved by not using the simplest stimuli. One might think that displays with a number of localized Gabor patches would suit the task. Gabor patches, a class of fundamental stimuli used in visual psychophysics, consist of a set of oriented bars generated by modulating a one-dimensional sine-wave intensity function by a two-dimensional Gaussian function. The resulting stimulus is a round patch with a limited number of dark and light stripes contained within it. However, Gabor patches are very similar to each other, and this becomes a problem in generating appropriate scenes. When only a few Gabor patches are used in each display, the visual input is too simple to provide the necessary environment for studying statistical learning. When more than a few Gabor patches are shown in the display, the visual system immediately invokes a number of previously developed midlevel representations based on configural or grouping mechanisms. For example, a cluster of horizontally oriented Gabor patches will pop out of a background array of randomly oriented Gabor patches. Thus, the statistical learning mechanism will work on these intermediate representations rather than on the structure controlled explicitly by the experimenter.

For our purposes, better control can be achieved when arbitrary configurations of complex, highly discriminable novel shapes are used as the elements. Because we are interested in the general properties of statistical learning, these properties can be investigated using complex shapes just as well as they can by using Gabor patches or other simple stimuli without interference from low-level mechanisms already in place. Moreover, by randomly assigning shapes across individuals, any effect of low-level feature similarity (or spontaneous labeling) can be eliminated. The key point is that eliminating low-level statistics requires the participants to rely only on the relevant higher level statistical relations that are embedded in spatial configurations of the shapes, and these statistics can be controlled very precisely. As a result, the task facing the learner is that any subelement in the scene can be associated with any number of others to form a new higher order feature. As opposed to Gaussian noise, this creates a problem of combinatorial noise; that is, even the bad combinations of subelements are potentially valid feature candidates. Thus, the observational learning paradigm allows us to investigate the general statistical rules for how mental representations of higher order features naturally emerge when the main challenge is not low visibility but unfamiliarity with the underlying structure of a complex input.

The Relation Between Embeddedness and Statistical Visual Learning

In a series of studies (Fiser & Aslin, 2001, 2002a, 2002b), we have shown that adults and infants can extract the conditional probability statistics between elements that cohere across scenes. The Fiser and Aslin (2001) study used an inventory of 12 complex shapes and introduced statistical structure by consistently pairing shapes in particular fixed spatial relations (base pairs) and generating a large number of scenes by presenting several base pairs (in all possible arrangements) within each scene (see Figure 1). The experimental designs used with adults are similar to those used in

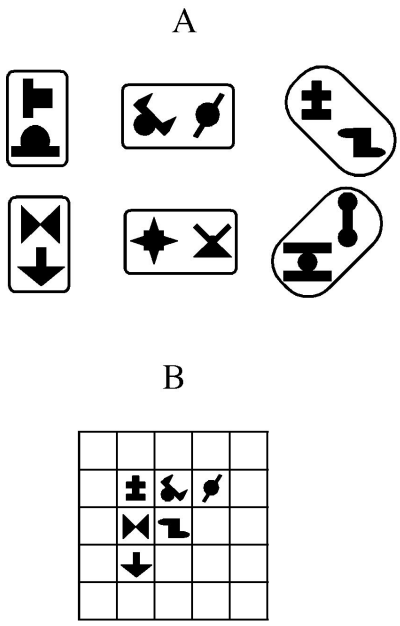

Figure 1. The stimuli used by Fiser and Aslin (2001) and in the current series of experiments. (A) Twelve arbitrary shapes, randomly assigned to pairs (or larger groupings), were used in all experiments. (B) A typical scene composed by three randomly selected base pairs.

the current series of studies and are detailed in Experiment 1. The Fiser and Aslin (2002b) study with 9-month-olds also used 12 shapes, but they were less complex and arranged in scenes containing fewer elements. The overall pattern of results from these two studies was clear and unequivocal. Both adults and infants exhibited a very strong sensitivity to the statistical relationships between elements after passive viewing of the visual displays for 2 to 20 min (depending on the experimental conditions). Even when joint probabilities did not differ between tested pairs of elements, and therefore individuals could not rely on the co-occurrence frequency of element pairs, they extracted conditional probabilities between elements in particular spatial configurations. Thus, both adults and infants are sensitive to the statistical structure of moderately complex scenes, and they extract this structure using conditional probabilities without the assistance of feedback during an observational learning phase.

The current series of experiments builds on the results of Fiser and Aslin (2001) by focusing on the question of feature hierarchy. Apart from the individual shapes, there is no obvious set of elementary building blocks in the groups of shapes in Fiser and Aslin (2001) that could serve as a universal inventory of features across all possible scenes. Rather, each scene can be decomposed into many different potential features varying in spatial scale and complexity. Individuals could remember individual shapes, shape pairs, triplets, and so forth, up to whole scenes. Moreover, these features are embedded within each other, forming a hierarchy from low-level subfeatures to high-level features. This situation is reminiscent of natural conditions in which the representation of scenes can be achieved in different ways, depending on which of the many embedded features are relevant. This embeddedness of features in the real world creates a very high level of redundancy among the potential features, and it is also the main reason why scenes encountered by human learners are confined to a very small part of the entire space of potential visual inputs. This high level of redundancy can, in principle, be exploited with an appropriate constraint to reduce the computational demand on the visual feature learning mechanism. Thus, our stimuli are suitable for exploring how humans remember and recognize visual inputs with highly embedded features and, in turn, can shed light on how humans learn new visual features. We used our experimental paradigm to investigate a potential constraint that the visual system might use to avoid the curse of dimensionality: eliminating (or reducing the weight of) features that are embedded in larger features but never appear outside these larger features. The following experiments examine whether the extraction of embedded features by humans complies with this embeddedness constraint.

Experiment 1

The goal of this experiment was to investigate how adult observers extract the relevant statistics of multielement scenes when the individual elements are arranged spatially into fixed triplets of shapes and these triplets are then used to create scenes. We used a modified version of the observational learning paradigm from Fiser and Aslin (2001), in which participants viewed a large number of six-element scenes composed of coherent base structures (in this case, two base triplets rather than three base pairs). These base triplets were presented in a grid such that they were spatially adjacent to each other. As a result, an element from one base triplet appeared most of the time adjacent to an element from a different base triplet, thereby forming one or more element pairs spanning two base triplets that had lower coherence (joint and conditional probability) than element pairs within a base triplet.

Because the structures in the scenes—the pairs, triplets, or *n*-tuples—themselves consist of distinctively shaped elements, each of which were decomposable into subparts, there was no a priori reason for the observers to select any individual shape or set of shapes as an important feature of the scene. For example, they did not know that the scenes had a triplet-based structure; they simply saw a grid containing six moderately complex shapes because each 2-s scene was presented during a passive learning phase. Thus, as far as the participants were concerned, each scene was full of complex embedded features. We consider a feature to be embedded in another feature if all of its elements are contained within the higher order feature. Thus, a coherent triplet of shapes, A-B-C, contains three embedded pairs (A-B, B-C, and A-C) and three single elements. At issue is how these various levels of the feature hierarchy are extracted during learning and represented in memory for later recognition.

Method

Participants. Undergraduates from the University of Rochester were participants in this experiment and in those that follow. They were paid $10 per session for their participation. Participants ranged in age from 18 to 25 years; the ratio of male to female participants was approximately 50:50 in all experiments. All participants were naive with respect to the purpose of the experiment and participated only in one experiment throughout this study to eliminate any cross-experiment contamination. Experiment 1 included 20 participants.

Stimuli. The same set of 12 arbitrary black shapes, of moderate complexity and presented on a white background, used by Fiser and Aslin (2001) were combined to generate 212 unique scenes by placing 6 of the 12 shapes in a 5 × 5 grid. The extent of the 5 × 5 grid was 11.4°, and the maximum size of each shape was 1.14°. The scenes were presented on a 21-in. Sony Trinitron 500PS monitor at 1,024 × 728 resolution from a 1-m viewing distance. Stimuli were controlled by a Macintosh G3 computer using Matlab and the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997).

Unknown to the participants, the 12 shapes were organized into four base triplets, in which each base triplet refers to three given shapes that always appeared in a particular spatial relation (Figure 2). Base triplets can be thought of as objects or rigid parts, in that if one of the shapes appeared in a given scene, the other two shapes always appeared in an invariant spatial relation to the first shape across all scenes during familiarization.

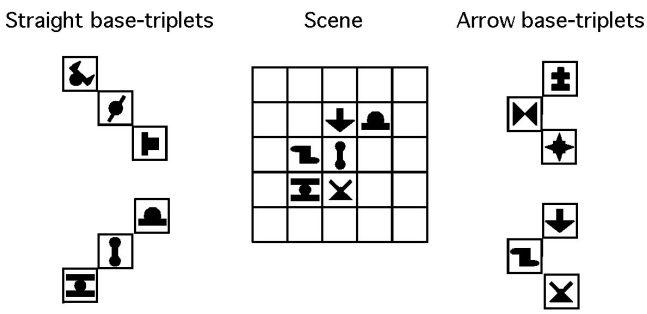

Figure 2. The base triplets of Experiment 1 and a typical scene.

The assignment of the 12 shapes to base triplets was randomized across participants to ensure that specific shape configurations were not unusual and more (or less) easily learned.

Spatial positioning of the elements within the visible grid eliminated uncertainty about positional coding, and spatial adjacency between base triplets ensured that the learning of base triplets was not facilitated by obvious segmentation cues. Given the constraints on base-triplet orientation and spatial adjacency, it was not possible to have a completely uniform distribution of individual shapes across all grid positions. However, within the central nine cells of the grid, the distribution was fairly homogeneous. Because all patterns appeared in the middle of the grid during the test phase, this mild inhomogeneity during the learning phase was unlikely to affect the outcome of the experiment.

A total of 112 different six-element scenes were generated by randomly positioning two of the four base triplets on the grid so that at least two elements of each triplet would have an element from the other triplet in a right or left neighboring cell of the grid. This arrangement ensured that the elements of the two triplets were completely meshed with each other and there was no way, based on midlevel grouping cues, to segment the two triplets. Each base triplet, and therefore each pair embedded in the triplet, as well as each of the 12 elements appeared an equal number of times across the 112 scenes. Shape pairs with neighboring elements that were contained entirely within a base triplet are referred to as *embedded pairs*. In addition to the designed triplet structures, a large number of accidental pairs, triplets, and higher order feature combinations occurred in the familiarization scenes.

Procedure. As in Fiser and Aslin (2001), there was a familiarization phase followed by a test phase. During the familiarization phase, participants saw each of the 112 possible scenes only twice (in random order) in an 11-min movie, with a scene duration of 2 s and a 1-s pause between scenes. Participants were told to pay attention to the continuous sequence of scenes, and no further instructions were given. There was a 3-min break between the familiarization phase and the next test phase.

After the familiarization phase, participants completed a series of temporal 2AFC tests with four types of trials in random order. The four test types consisted of single elements, pairs, triplets, and quadruples, so that the participants would not pay special attention to either the pair or triplet test trials or to shape combinations that were the focus of our investigation. In single trials, two grids appeared after each other, each containing a single element in the middle of the grid; in pair trials, an isolated pair appeared in the center of each grid; and so on. In the single trials, the participants were asked to choose which of the single elements they thought appeared more frequently during familiarization. In all the other types of test trials, the participants were asked to select the structure that looked more familiar based on what they viewed during the familiarization phase. Because all single shapes appeared an equal number of times during familiarization, the single trials were filler trials with no correct answer. Similarly, because there were no strong quadruple structures embedded in the familiarization phase, both quadruples in the quad trials were set up with dummy shape combinations never seen before together to serve as foils. In other words, in all quadruple trials, both scenes were made up from four elements that never appeared in that particular configuration, and so there was no correct answer as to which would be more familiar. Similarly, there were dummy pair and triplet trials; both structures in the trial had element arrangements that never appeared during familiarization. These trials were added to keep the appearance frequency of individual shapes in each type of test trial equal. All single shapes and configurations of shapes were presented centrally during the test trials, and the order of the two test items in a trial was counterbalanced. All together, there were 38 test trials, including the 12 nondummy trials. Each test display was presented for 2 s with a 1-s pause between them. Participants had to press a computer key (the 1 or 2 key) depending on which of the two test patterns was judged to be more familiar (or more frequent in the single-element test trials). The presentation order of the test trials was individually randomized.

Intermixed with the foil trials, the 12 key test trials consisted of two types: (a) comparison of a base triplet with a triplet composed of shapes that never appeared in that particular spatial configuration (pairwise or triplet), and (b) comparison of an embedded pair from within a base triplet with a pair of shapes that never appeared next to each other during the familiarization phase but could appear in the same familiarization display. In both of these types of test trials, one pair or triplet was statistically coherent (conditional probabilities [CPs] = 1.0) and the foil pair or triplet was incoherent (CPs = 0.0). We were particularly interested in whether participants first extracted the pair structures embedded in the triplets or whether they learned the larger triplet structures and the smaller embedded structures in parallel.

Results and Discussion

Because in the test trials with single elements or quads there was no correct answer (neither member of the trial had higher statistical coherence), the participants were expected to select each single shape or quad on a test trial as often as the other member of the trial. Indeed, in this experiment, as well as in the tests using dummy trials in all subsequent experiments, the participants showed no significant deviation from chance performance ($p > .05$). The results of the crucial tests in Experiment 1 with pairs and triplets are shown in Figure 3. After 11 min of viewing during the familiarization phase, participants reliably selected the base triplets over random triplets, $t(19) = 2.89$, $p < .001$. (All t tests in the current study were two-tailed.) Because in this experiment joint and conditional probabilities covaried, either of these two sources of statistical information could have been used to select the base triplets over the random triplets. In contrast to the triplets, participants did not reliably select the embedded pairs over random pairs, $t(19) = 0.81$, $p = .428$. In other words, although the base triplets were reliably distinguished from random triplets, the coherent subfeatures of the base triplet (i.e., the embedded pairs) were not discriminable from a random pairing of elements.

However, because the mean performance on the pair test trials was slightly above 50%, the difference in performance between pairs and triplets was not significant, $t(19) = 1.42$, $p = .171$. To determine whether this absence of a significant difference was due to statistical power, we collected data from an additional 20 individuals and combined the results from all 40 participants. The mean percentage of correct responses for pairs and triplets remained almost exactly the same (56.1% and 67.9%, respectively). The preference for base pairs still failed to exceed chance, $t(39) = 1.52$, $p = .138$, and the preference for base triplets remained significantly above chance, $t(39) = 3.83$, $p < .0005$. However, now with more statistical power, the difference in performances between pairs and triplets approached significance, $t(39) = 1.83$, $p = .074$.

When the participants were asked after the test whether they noticed anything during the practice, typically they did not report the true underlying structure of the scenes. They also reported that they were guessing during the test, and their performance and their confidence in their performance did not correlate well. Even though these observations were reported during an informal debriefing chat and they were not quantified rigorously, they suggest that the knowledge of the underlying triplet structure of the scenes emerged through an implicit learning process.

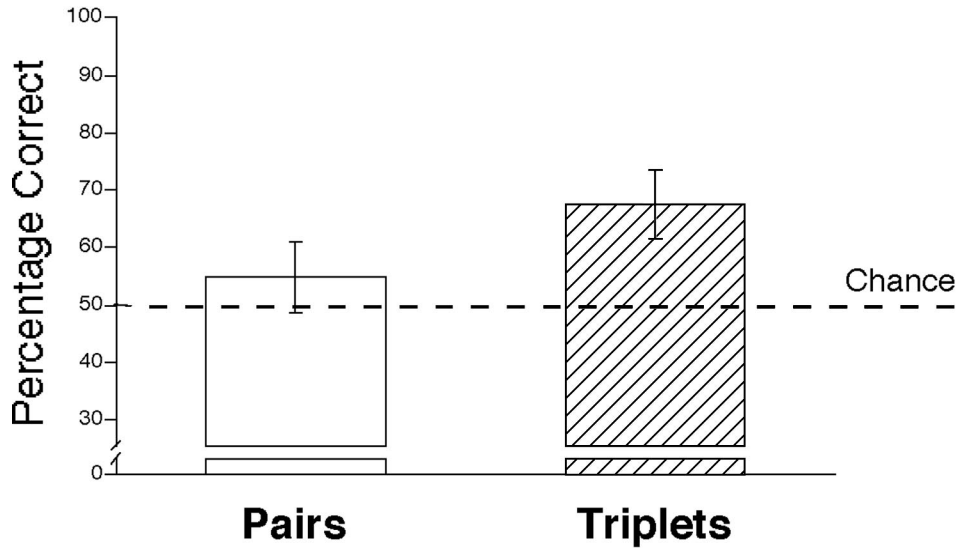

Figure 3. The results of Experiment 1. Participants chose base triplets reliably more often over random triplets, but they failed to do so with base pairs that were embedded in the triplets over random pairs. Error bars represent ± 1 *SEM.*

These results appear to contradict the notion that a larger complex feature (the base triplet) is built up hierarchically from its constituents. This may seem paradoxical because it is clear that the representation of the base triplets that enabled the individual to select them over random triplets must in some way use the information carried by its embedded pair-based structure. However, on further consideration, there is no paradox if one assumes that these results only imply a reduction in the accessibility of the substructures embedded in larger, more complex structures once these larger structures have been learned. In other words, the representation of the larger triplet structure relies on some information carried by the pair-based substructures, but the substructures are not represented explicitly (at least not all of them) so that they can be accessed as separable features. Note that by the term *represented explicitly* we refer to the fact that these elements are more accessible for recall and, therefore, presumably for using them as building blocks for higher order features. We do not postulate anything about the neural realization of explicitness in terms of representational units, zero storage for all embedded elements, or computational weighting to favor the larger structures while suppressing the smaller ones. For example, the actual realization could weight positive and negative evidence from subsets of the input, and if the sum passes some threshold it is taken as evidence for the existence of the higher order feature in the scene. The difference between the triplets and the embedded pairs in this scheme is that the triplets gain evidence that passes the threshold while the pairs do not. Thus, the triplets are represented strongly enough to be building blocks of further memory traces while the pairs are not (or less so), even though they might activate some limited trace.

It is important to note that the discrepancy in performance between triplet and pair test trials is not because the triplet tests were easier than the pair tests as a result of the fewer number of potential triplets compared with potential pairs in the scenes. The numbers of possible pairs and possible triplets (considering only neighboring elements) across all scenes of the familiarization movie were 785 and 845, respectively. This means that the average number of potential pairs and potential triplets in each familiarization scene was 7.00 and 7.54, suggesting that a particular triplet should certainly be no more probable to remember than a particular pair. The pair–triplet difference in performance also cannot be explained by special Gestalt principles of shape arrangement. Although it is true that two of the shapes used in the experiment, referred to as arrow and line, could represent an easy-to-encode special case, a breakdown of the participants' errors revealed no differential bias to select either of these shapes (30% vs. 35% for lines and arrows, respectively).

Nevertheless, two possible explanations of the results should be considered before accepting the hypothesis that representations of substructures embedded within a larger structure tend to be suppressed during visual learning. The first possible explanation of the results is based on the method of testing, and the second is based on the structure of the familiarization movie. In particular, the first potential explanation is that our test method of using mixed test trials with single elements as well as large structures up to quadruples interfered more strongly with remembering the smaller pair features than with remembering the larger triplet features. In other words, across all test trials, many more pairs (all the pairs and the embedded pairs of triplets and quadruples) than triplets were shown, and this could lead to potentially more confusion regarding remem-

bered pairs from the familiarization phase. The second possible explanation is a scale effect; that is, when the familiarization scenes are composed exclusively of larger triplet structures and the scenes always contain many elements, this might lead participants to unconsciously attend to larger structures. As a result, processing of the smaller pair structures would be suppressed in general rather than because they are embedded in the triplets. The next two experiments tested these possibilities.

Experiment 2

Experiment 2 tested whether participants could extract pair-based structures from scenes that were, in all respects other than their underlying generative structure, identical to those used in Experiment 1. If they could, then it would rule out the hypothesis that the failure in Experiment 1 to extract embedded pairs was merely the result of a higher interference of the test items with memory traces of shape pairs.

Method

Participants. Twenty new naive undergraduates from the University of Rochester participated in Experiment 1.

Stimuli, design, and procedure. Experiment 2 was identical to Experiment 1 in all respects except for the structure of the familiarization scenes and the specific pair and triplet tests. In Experiment 2, six base pairs were used instead of four base triplets for generating the scenes. Each scene was based on one scene of Experiment 1, so that the empty and the filled grids in the scene used in Experiment 2, and in the corresponding scene in Experiment 1, were the same. In other words, the silhouette of the six-shape scenes defined by the filled grid positions was identical to the original scenes in Experiment 1. However, the scenes of Experiment 2 were composed of three of the six base pairs rather than two of the triplets as in Experiment 1. Thus, for a naive observer, the scenes appeared to be exactly the same in the two experiments; only the higher order structures were different. Each base pair and each single element appeared the same number of times across the scenes presented during familiarization.

The critical test trials in Experiment 2 consisted of pairs of shapes because the coherent structure in the scenes was pair based. The pair test trials had exactly the same silhouette as in Experiment 1 but used the base pairs of Experiment 2 rather than random or embedded pairs. The single-shape and quadruple test trials were identical in Experiments 1 and 2 and served as foils to mask our interest in the pair test trials. The triplet test trials, which were relevant in Experiment 1, were no longer relevant in Experiment 2 and had the same silhouette as the triplet trials in Experiment 1 but they were foil trials with random structures.

Results and Discussion

Participants in all trials with dummy test elements showed chance performance. Figure 4 shows the pair results of Experiment 2. Although the number of potential pairs and triplets and their exact absolute and relative positions during practice were identical to those in Experiment 1, the participants had no difficulty choosing the base pairs over the random pairs, $t(19) = 4.25$, $p < .0005$, under the same test conditions as in Experiment 1. This finding excludes the possibility that the results of Experiment 1 were due to a higher interference of the test items with memory traces of shape pairs.

Experiment 3

Experiment 3 served to further investigate the results from Experiment 1. Although Experiment 2 showed that participants could learn a pair-based structure when it was the only statistical structure embedded in the six-element scenes, it is possible that scenes containing two different levels of statistical structure might prevent individuals from extracting both. Such a possibility would explain why participants in Experiment 1 showed a difference in their ability to extract triplet-based structures versus pair-based structures, independent of the question of embeddedness. Experiment 3 was designed to determine whether participants were able

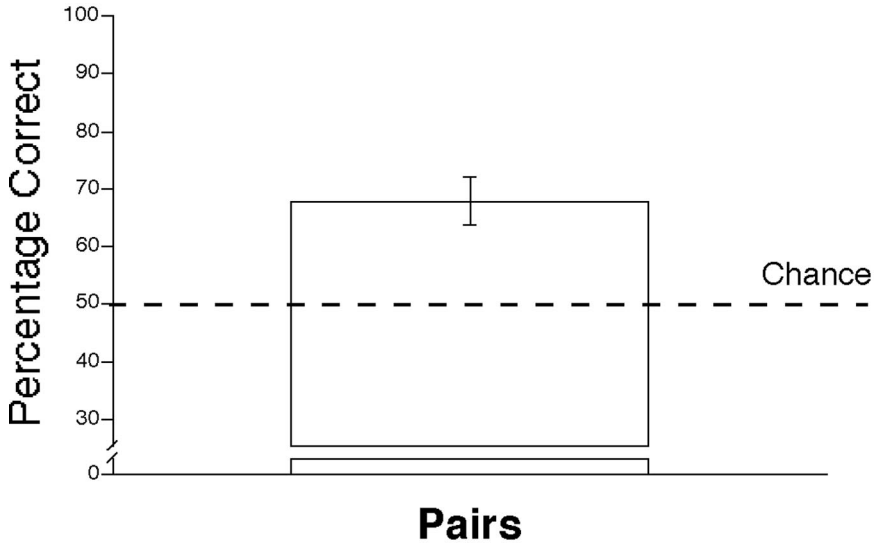

Figure 4. The result of Experiment 2. Participants preferred base pairs over random pairs. Error bars represent ± 1 *SEM*.

to extract both pair-based and triplet-based structures in parallel when the tested pairs were not parts of the tested triplets.

Method

Participants. Twenty new naive undergraduates from the University of Rochester participated in Experiment 3.

Stimuli, design, and procedure. Experiment 3 was identical to Experiment 1 in all respects except that the structure of the familiarization scenes contained not only base triplets but also nonembedded base pairs. The 12 shapes were grouped into three base pairs and two base triplets. Half of the familiarization scenes consisted of the two base triplets, and the other half consisted of the three base pairs. The silhouettes of the scenes as well as the silhouettes of the triplets and pairs were the same as in the previous two experiments, and each triplet, pair, and single shape appeared an equal number of times during the familiarization phase. During familiarization, the triplet- and pair-based scenes were intermixed and presented in random order, and participants were given no indication as to what type of scene was being presented.

The dummy single shape and quadruple test trials were identical to those in Experiments 1 and 2. The pair and triplet tests of Experiment 3 had exactly the same silhouette as in Experiments 1 and 2, but the actual shapes were the base pairs of the current experiment compared with random pairs and the base triplets of the current experiment compared with random triplets.

Results and Discussion

Once again, participants in all trials with dummy test elements showed no significant deviation from chance performance. Figure 5 shows the pair and triplet results of Experiment 3. Both base pairs and base triplets were chosen significantly more often than the random pairs and random triplets, $t(19) = 4.22$, $p < .0005$, and $t(19 = 3.47$, $p < .005$, for pairs and triplets, respectively. These results demonstrate that when the most significant structures within a scene consist of both triplets and pairs, participants have no difficulty becoming sensitive to features of different complexities (quantified by the number of elements included) in parallel. These findings rule out the possibility that participants in Experiment 1 failed on the pair-based test trials because of a scale effect (i.e., because they could not simultaneously extract both pair-based and triplet-based structures from the scenes).

Experiment 4

The results of Experiments 2 and 3 excluded the simplest explanations of the findings of Experiment 1 based on generic biases, either in the test phase or in the familiarization phase, to attend to triplets over pairs. Thus, the most parsimonious interpretation of the results of Experiment 1 is that when a representation of a complex visual feature (the base triplet) that contains a number of embedded simpler features (any pair within the triplet) is being formed during observational learning, separate representations of the embedded features are suppressed. A corollary of this interpretation is that whenever two features of the same complexity are present in a scene, the one embedded in a larger feature must be represented to a lesser degree than the other, which is not included in a larger feature. Experiment 4 tested this corollary directly.

Method

Participants. Twenty naive undergraduates from the University of Rochester, who were paid $10 per session for their participation, served as participants.

Stimuli and design. The same 12 arbitrary complex black shapes on a white background and the same 5 × 5 grid were used in Experiment 4 as in all the previous experiments. Unknown to the participants, the 12 shapes were organized into two base quadruples, or base quads, and two base pairs (see Figure 6). As before, the specific assignment of the 12 shapes to base quads and base pairs was randomized across participants.

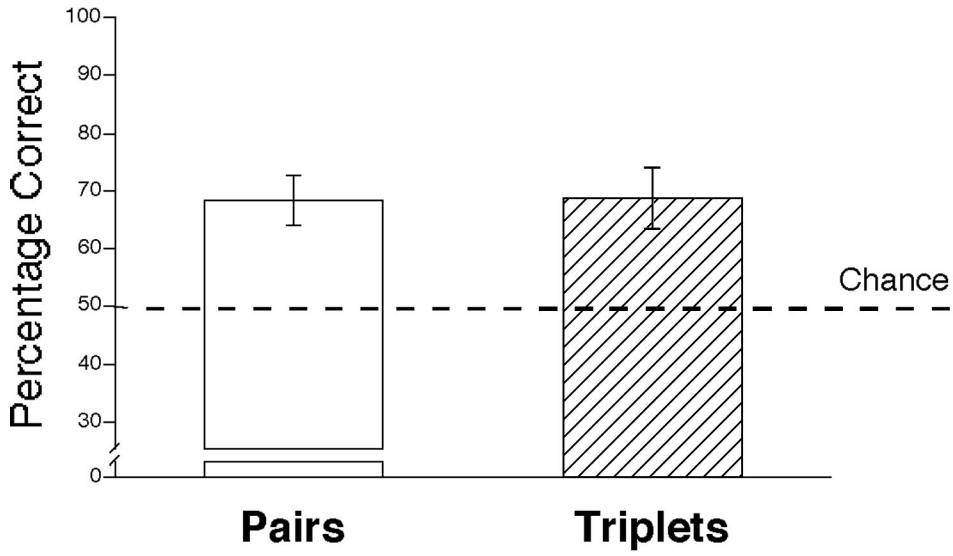

Figure 5. The results of Experiment 3. Participants could reliably select base pairs and base triplets over random shape combinations when scenes composed from only triplets or only pairs were randomly intermixed during the familiarization phase. Error bars represent ± 1 *SEM*.

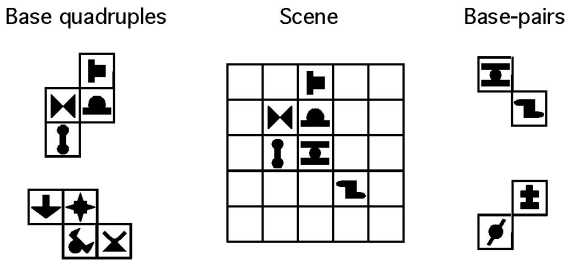

Figure 6. The base quadruples and base pairs of Experiment 4 and a typical scene.

A total of 120 different scenes were generated by randomly positioning one of the two base pairs and one of the two base quads on the grid so that at least two elements of the quad would have an element from the base pair in a right or left neighboring cell of the grid. Each shape of the 12 elements, each base pair, and each base quad appeared an equal number of times across the 120 scenes. Just as in Experiment 1, shape pairs with neighboring elements that were contained in a base quad are referred to as *embedded pairs*.

Procedure. The familiarization phase in Experiment 4 was divided into two parts, each part followed by a test. During each familiarization phase, participants saw each of the 120 possible scenes only once (in random order) in a 6-min movie, with a scene duration of 2 s and a 1-s pause between scenes. Similar to all previous experiments, participants were not given any specific task other than paying attention to the scenes. There was a 3-min break between each familiarization phase and its corresponding test phase.

After each of the two familiarization phases, participants completed a temporal 2AFC test phase, which was slightly different from the tests in the previous experiments. During the test, participants saw 26 trials, which were only pair trials or quadruple trials in random order. In the quad trials, base quads (one of the two clusters on the left side in Figure 6) were compared with a random arrangement of four shapes. In the pair trials, base pairs were compared with pairs of shapes that never appeared next to each other during the familiarization phase. There were two types of base pairs compared with random pairs. The first type consisted of pairs embedded within the base quads (i.e., two neighboring shapes from a base quadruple in Figure 6). The second type consisted of the base pairs of the scenes that were independent from the base quads and were not embedded in any higher order structure (i.e., one of the two pairs on the right column in Figure 6). Because the base pairs in this experiment were always diagonally arranged for better mixing with the quadruples, all test pairs, coherent or random, were arranged diagonally, and the main directions of the diagonals were counterbalanced. Each test pair or quadruple was positioned in the middle of the grid. After completing the first test phase, the participants were exposed to the second half of the familiarization phase and then completed the same temporal 2AFC test phase a second time. The order of the test trials was randomized individually in the first and the second test sessions.

Results and Discussion

Figure 7 shows the results of Experiment 4. The pattern of results in the two test phases was very similar after 6 and 12 min of familiarization, demonstrating that the test phase after the first round of practice did not bias the participants to attend more to the statistical structures of interest (the pairs and the quads). Confirming this, a 2 (round) × 3 (test type) analysis of variance (ANOVA) revealed no main effect for round, $F(1, 19) = 0.58$, $p = .453$, $\eta^2 = 0.03$. In contrast, there was a main effect of test type, $F(2, 38) = 11.82$, $p < .0001$, $\eta^2 = 0.38$, with no interaction, $F(2, 38) = 0.33$, $p = .717$, $\eta^2 = 0.02$. Even after only 6 min of passive viewing, participants showed a significant preference for extracting the familiar structure of the base quads over random quads and the nonembedded pairs over random pairs, $t(19) = 4.16$, $p < .0005$, and $t(19 = 3.56$, $p < .005$, for quads and nonembedded pairs, respectively. However, participants failed to distinguish the em-

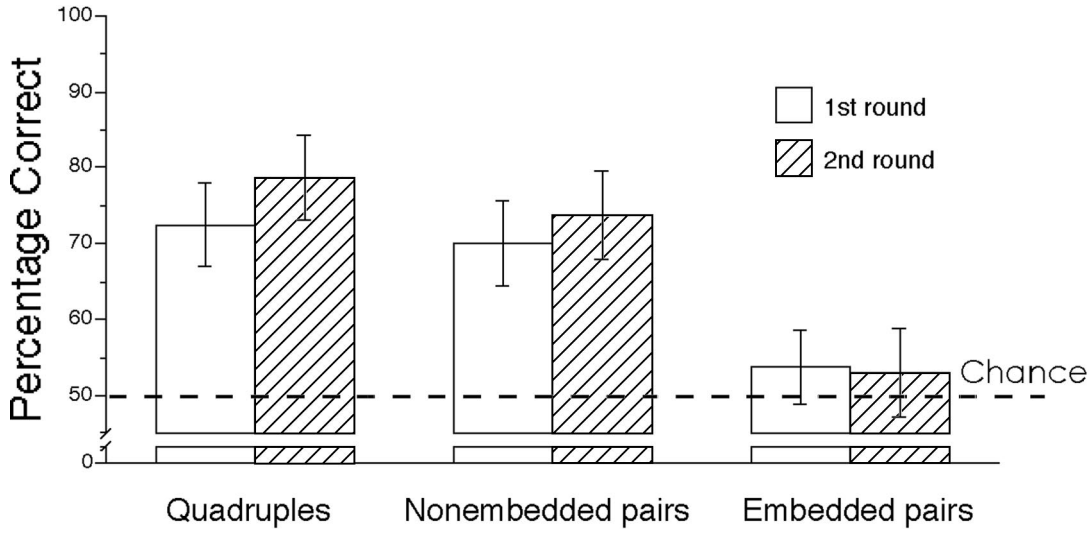

Figure 7. The results of Experiment 4. Both base quadruples and base pairs were preferred significantly over random combinations of elements. However, participants were unable to distinguish pairs embedded in the quadruples over random pairs. This pattern of results did not change after doubling the duration of the familiarization phase. Error bars represent ± 1 *SEM*.

bedded pairs from random pairs, $t(19) = 0.78$, $p = .444$, even though these embedded pairs were seen as often as the nonembedded pairs and the quads that included these embedded pairs were easily discriminated from random quads. The difference in performance between the quads and the pairs embedded in the quads, as well as between the embedded pairs and the base pairs, was significant, $t(19) = 2.97$, $p < .01$, and $t(19 = 2.31$, $p < .05$, for quads versus embedded pairs and for nonembedded versus embedded pairs, respectively.

An additional 6 min of familiarization improved the participants' performance, albeit not significantly, on both the tests of base quads (5.25% increase) and base pairs (3.75% increase). However, additional familiarization did not raise the participants' performance for the embedded pairs over random pairs to above-chance levels, and in fact it decreased the participants' performance slightly (−0.625%). These findings confirm the initial conclusion drawn from Experiment 1 with pairs embedded in triplets and supported by the controls in Experiments 2 and 3. That is, the representations of complex features that develop implicitly during passive observational learning are biased to highlight higher order features and to suppress embedded features that are a part of and redundant with a larger, more complex feature.

Experiment 5

In the four preceding embedded experiments, we found an interaction between how well features are learned and how they are linked together by being present in the same coherent visual structure (what we refer to as a *chunk*). The representation of features embedded in more complex higher order features was suppressed, as demonstrated by participants' apparent inability to discriminate them from random structures in 2AFC tasks. However, it is not clear what determines the size and boundaries of a chunk. In the preceding embedded experiments, element co-occurrence and predictability always covaried. Either or both of these two statistics could affect the process of chunking.

In previous research with multielement scenes (Fiser & Aslin, 2001), we showed that when element co-occurrence (relative frequency) and element predictability (conditional probability) covary, individuals tend to remember element combinations (higher order features) with high co-occurrence and high predictability. The same study also demonstrated that when element co-occurrence does not differ between two pairs, individuals can rely on the conditional probabilities (predictability) of pairs exclusively, and they preferentially encode pairs with high predictability. In these studies, when element co-occurrence of the tested pairs was equated, the tested features with low and high predictability had never appeared in the same scene during learning. Thus, individuals collected information independently during the familiarization phase about the two types of features that they were asked to compare in the test phase.

In Experiment 5, participants viewed scenes with different embedded structures that had different predictability regardless of their co-occurrence frequency. These scenes approximate real-world situations better than scenes in the previous embedded experiments. Because feature co-occurrence can be separated from predictability only by adjusting the appearance frequency of certain individual elements within the feature, the experimental question that can be explored with such scenes is as follows: What determines the manner in which complex visual input is chunked into new features during the observational learning process? Is it the appearance frequency of the individual elements in the scene (i.e., how often we see them), or their predictive power (i.e., how much we can rely on them in predicting the appearance of some other elements) that constrains statistical learning? Experiment 5 investigated this question.

Method

Participants. Thirty naive undergraduates from the University of Rochester, who were paid $10 per session for their participation, served as the participants.

Stimuli and design. The same 12 arbitrary complex black shapes on a white background and the same 5 × 5 grid were used in Experiment 5 as in all the previous experiments. Unknown to the participants, the 12 shapes were organized into two base sextuples (see Figure 8). As before, the specific assignment of the 12 shapes to the base sextuples was randomized across participants. In addition to the coherent sextuple structures, two shapes from each sextuple were selected to serve as an additional noise element in creating seven-element scenes when combined with the other

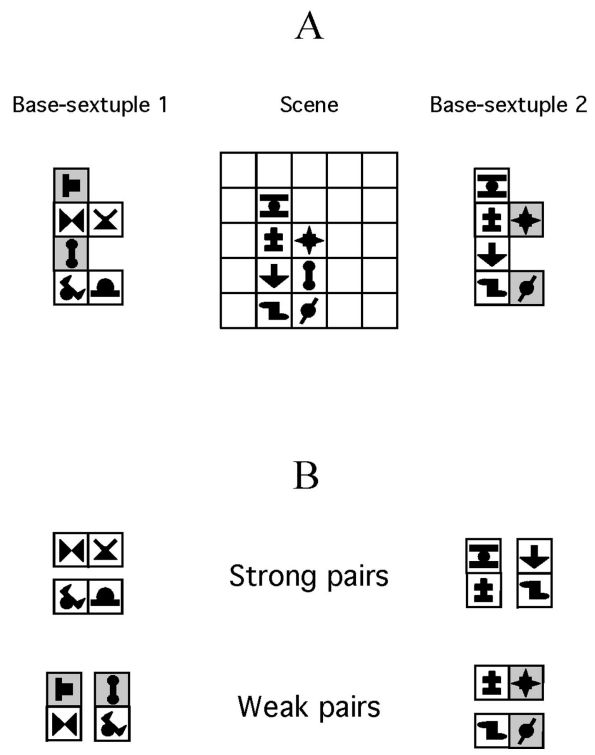

Figure 8. The stimuli of Experiment 5. (A) The scenes of Experiment 5 were generated by using one of the sextuples and one noise element from the other sextuple. The two possible noise elements in each sextuple are marked with a gray background for illustrative purposes only. (B) Because of the choice of the sextuples' outline shape and the noise elements, both large structures (sextuples) contained strong pairs with high predictability and low appearance frequency and weak pairs with low predictability and high appearance frequency. The orientation and position of these elements were completely balanced across the scenes.

sextuple. A total of 184 different scenes were composed of seven shape elements by putting one of the sextuples in random position onto the 5 × 5 grid and adding one of the two noise elements from the other sextuple so that the noise element was adjacent to at least one element of the sextuple (see Figure 8).

This method of scene generation resulted in the four noise elements appearing 1.5 times as often as all the other shapes across the entire set of scenes. In addition, all the pair-based co-occurrence frequencies, and therefore all the joint probabilities, within each sextuple were identical. Shapes within pairs of the sextuples that did not involve any noise elements, referred to as *strong pairs*, had higher (in fact, perfect) predictability (CPs = 1.0), because one element of the pair predicted perfectly the appearance of the other element. In contrast, pairs of shapes within the sextuples that did contain noise elements, called *weak pairs,* had lower (CPs = 0.66) predictability, because the noise elements could appear in different pairings one third of the time. Thus, strong pairs had high predictability but lower element appearance, whereas weak pairs had lower predictability but higher element appearance (see Figure 8B). Therefore, any difference in participants' performance on a test comparing strong and weak pairs would answer the question of whether predictability or appearance frequency determines chunking. The selection of the noise elements was such that horizontal and vertical arrangements and relative position of the tested strong and weak pairs within the sextuples were balanced.

Procedure. During familiarization, participants saw each of the 184 possible scenes only once, in random order, in a 9-min movie, with a scene duration of 2 s and a 1-s pause between scenes. Participants were not given any specific task, and there was a 3-min break between the familiarization and the test phase.

After the familiarization phase, participants completed a temporal 2AFC test phase very similar to the previous embedded tests, with trials using single elements, pairs, and quadruples but not triplets. In Experiment 5, there were no dummy trials without a correct answer. In the single trials, the four high-frequency noise elements were compared with other shapes that had appeared with lower frequency. In the quadruple trials, embedded quadruples of the two sextuples were compared with random quadruples. The key test trials were the ones in which a strong or a weak pair was compared with a random shape pair, consisting of shapes that never appeared next to each other during the familiarization phase. Because the main question of interest was the difference between weak and strong pairs, the test phase was split into two halves: First, the trials with pairs were run, and then the trials with quadruples and single elements were conducted.

Results and Discussion

Although scenes composed of a sextuple and a single noise element might seem to be easily decomposed into its two constituents, most of the participants did not report during the posttest debriefing that they had became aware of the true structure of the scenes during the experiment. Figure 9 shows the results of Experiment 5. A one-way ANOVA with four levels revealed a main effect of test type, $F(3, 87) = 5.63$, $p < .0015$, $\eta^2 = 0.16$. In all four types of tests, participants' performance deviated significantly from chance level, $t(29) = 2.73$, $p < .011$, $t(29) = 6.84$, $p < .0001$, $t(29) = 4.43$, $p < .0001$, $t(29) = 5.64$, $p < .0001$, for singles, strong pairs, weak pairs, and quads, respectively. That is, participants performed significantly above chance in identifying which of the individual shapes, and also which pair or quad structures, were shown more frequently during the familiarization phase. However, there were significant differences in performance between the different types of test stimuli. Strong pairs and quadruples embedded in the sextuples were identified correctly equally well, $t(29) = 0.32$, $p = .751$. In contrast, weak pairs were identified significantly less often than either strong pairs or quadruples, $t(29) = 2.25$, $p < .05$, $t(29 = 2.50$, $p < .05$, for strong versus weak pairs and quadruples versus weak pairs, respectively. This result shows that, although weak and strong pairs appeared the same number of times, both in the same positions within the grids and within the larger structure of sextuples, and although elements of the weak pairs appeared more often than elements of the strong pairs, strong pairs were encoded better and remembered more because of the higher predictability of their elements. A corollary of this finding is that weak pairs serve as the break points for dividing a large coherent structure into chunks.

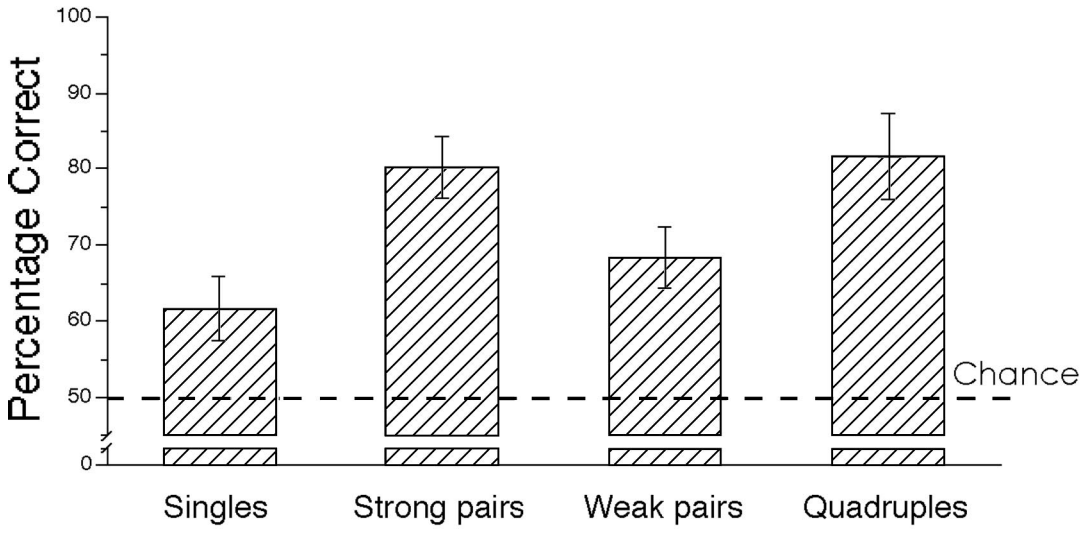

Figure 9. The results of Experiment 5. In all four tests, participants chose significantly more often the base structures over random structures or over the single element with the higher appearance frequency. However, the weak pairs were selected significantly less frequently than the strong pairs. Error bars represent ± 1 *SEM.*

General Discussion

In five experiments, we investigated a key constraint, *embeddedness,* that could enable human adults to extract the statistics of visual information to form memories of new higher order features from complex scenes without suffering from the curse of dimensionality. Overall, we found that, although humans rely heavily on visual statistics, including both joint and conditional probabilities among feature elements, their performance deviates in a number of ways from simply extracting different image statistics independently and forming representations based on the most prominent statistics in the set of scenes.

We found that humans extract independent parts—highly coherent subsets of elements—of different levels of complexity from the scene in parallel and without an explicit task (Experiments 1–3). However, parts of the same complexity were remembered to a different degree depending on whether or not they were embedded in a larger, more complex cluster of elements: Features embedded in larger features were represented to a lesser degree (Experiment 4). This implies that not all the embedded features that are parts of a larger whole are explicitly represented once a representation of the whole has been consolidated. We call this phenomenon the embeddedness constraint of statistical learning. In the final experiment, we found that embedded parts with strong predictability are encoded significantly better than embedded parts with less predictability, even if the elements of the parts with less predictability appeared more often across all the scenes presented during the familiarization phase. A consequence of this finding, we argue, is that when a large, complex feature consists of elements with unequal predictive power, the representation of the large feature is broken into parts, or chunks, at the "fault lines" between elements that have low predictability with each other. We propose that this is a statistical definition of chunking that is consistent with and complements the classical definition that is based on the functional limits of short-term memory (Chase & Simon, 1973; Ericsson, Chase, & Faloon, 1980).

Statistical Learning With the Constraint of Embeddedness

Although our use of a 2AFC posttest did not allow us to directly reveal the process of learning, we propose that our postfamiliarization results are indicative of how the visual system develops new higher order features of initially unfamiliar inputs. We suggest that the embeddedness constraint reduces the overall number of features that are required to represent the full range of structures in scenes by reducing the complexity of the representation of the large coherent structures. By analogy, consider an example from the domain of text with the task of identifying words that are built up from lower level embedded features of letters, letter pairs, and letter combinations of different length. For the sake of clarity, we allow for only two types of features: (a) adjacent letter pairs and (b) nonadjacent letter pairs with a "wild card" between them that can be any letter. Assume that the new word "redundancy" appears for the first time in the written corpus. In our recognition system, there are 17 potential pair features embedded in the larger structure of the word *redundancy* that would be activated by the word, and all these features should be considered for representing the word and to learn any new higher order features that involve the word. Nevertheless, to represent the unique letter string *redundancy* in a word-recognition task involving a large number of other words, only four of these pairs may be sufficient (e.g., *re, d_n, a_c, cy*). As new words are added to the corpus of text, some of the other 13 features might be included in a bootstrapping manner to represent *redundancy* as a unique string because of the need for greater specificity to distinguish it from other words. However, if there are no other words in the corpus that also happen to contain these four pairs, then this small feature set may suffice to provide a unique representation of the word *redundancy* for all the purposes of the system. Also, any additional learning of new higher level representations will act only on this limited sufficient set of features rather than on the entire range of all possible features of the word. The key point is that, in general, the vast majority of the potential features (13 of the 17) are not used to represent this higher order structure (the word *redundancy*), and because the limited representation is sufficient for all the required tasks, it may not even be evident to the representational system that the structure of the word is only partially specified. Consequently, if there is a reliable mechanism that is biased to represent the largest chunks in the input in a minimally sufficient manner, rather than using a full representation of all possible features, this constraint can eliminate the curse of dimensionality.

Suppressing the internal representation of all unnecessary or excess features is not only advantageous from the point of view of reducing complexity, but it is also naturally implemented in an adaptive, competitive neural network that is sensitive to statistical regularities. For example, neural network simulations that learn new higher order complex features from multielement scenes by competitive learning using conditional probability statistics inherently suppress the embedded structures while developing representations of the largest coherent parts in the scene (Dayan & Zemel, 1995). Similar competitive learning mechanisms were hypothesized to underlie visual feature formation in the cortex (Bienenstock, Cooper, & Munro, 1982).

How should the minimally sufficient set of features of a new, complex visual structure be selected? On the basis of the results of Experiment 5, we suggest a mechanism by which the visual system seems to achieve this less redundant representation. According to this mechanism, when the visual system faces complex scenes composed of elements with uneven appearance frequencies and predictive power, it extracts new complex features based on a combination of preexisting partial representations with the highest predictive power (high CPs). The largest combinations of elements with high predictability serve as coherent chunks, each of which forms a single, new higher level feature, whereas combinations of elements with low predictability serve as natural breakpoints to separate the whole scene into its parts. Returning to our analogy with the word *redundancy,* if the previously learned corpus of words made it evident that the *_ncy* pattern is a highly typical letter combination across many words, but the letter before the "n" is highly variable, then the system would break the representation before the letter *n* and use the single feature of *n_y* to represent the chunk, thereby neglecting the explicit representation of both the *nc* and the *cy* embedded features.

It is important to emphasize that, although this scheme reduces redundancy, it is very different from the models of early vision, in which redundancy reduction is pursued without allowing the loss of input information even when additional constraints, such as sparseness, are applied (Bell & Sejnowski, 1997; Olshausen & Field, 1996). Quite the contrary, the chunking model deliberately neglects input information by selectively coding only a part of the incoming information while searching for the minimal code that is sufficient for interpreting the input. Naturally, this scheme requires an incremental bootstrapping mechanism of feature learning, so that when the minimal code proves insufficient it could get expanded.

Because the input to the statistical learning mechanism is now constrained by the predictability power of lower level descriptors, when predictability is fairly uniform across elements, the emerging new higher order features will generate a holistic or view-based representation such that the features used to represent the scene are not restricted to a subregion of the scene. On the other hand, when predictability varies considerably across subregions of the scene, thereby creating a number of breakpoints, a structural description or part-based set of new complex features is more likely to develop. In the same way, predictability can be the basis of the emergence of individual objects from the scene (Fiser & Aslin, 2001).

How would the visual system know the predictive power of any subfeature within a new potential feature in an unknown scene to select the ones with the highest power to form chunks and the minimally sufficient set of features? This is where incrementality of the learning process becomes important. Because the initial representation of the unknown scene uses previously developed features, the predictability of those subfeatures must have been assessed earlier during the learning process in some different context. The simplest scheme is to rely, at least initially, on those subfeatures with the highest predictive power in the previous context to form new higher level representations. These highly predictive subfeatures are not determined by how prominent they are in the scene (i.e., by their relative frequency of occurrence) but by how much predictive power they have for describing other elements within the scene or the scene itself and how easy they are to extract (Jacobs, 2003). Specific examples of such prominent subfeatures include not only those that are given by built-in or early developed midlevel mechanisms (e.g., luminance and chromatic contrast, T- and Y-junctions, contour proximity, and other Gestalt cues; see Pomerantz & Kubovy, 1986) but also more complex subfeatures that are combined to form higher level features by a constrained statistical learning process.

However, most of the subfeatures clearly do not have one single general predictability value that is relevant in all cases. Take, for example, the standard and the fold-out versions of mobile phones. In the standard case, the shape, position, and lighting conditions of the keypad section are highly predictive of the same attributes of the display section; thus, a single chunk could be used to represent both sections, whereas with the newer fold-out phones this is not the case. How can our scheme handle the fact that the same feature in different contexts might have very different predictabilities? Initially, the subfeature within the wrong context will still be used for forming higher order features (e.g., the overall shape of the phone when both the display and the keypad are visible) according to their assumed predictability. However, when the resulting higher order features turn out not to be useful (e.g., during a visual search for the phone in a cluttered environment), the final representation will eventually reject (or uncouple) the features based on the subfeature with weak predictive power, and the representation of the subfeature will be suppressed within this context. Therefore, a subfeature can be highly significant and heavily used in one context and completely absent from the representation in a different context.

The Relation Between Natural Scene Perception and Observing Multielement Displays

As described early in this article, we purposely created displays that lack real-world features beyond the individual shapes themselves, acknowledging that the adult visual system already has a large built-in and developed set of detectors for extracting dependencies among pixels and features of various complexity. By choosing our stimuli, we tried to control these dependencies in two ways. First, we acknowledge that lower level dependencies operating at the level of the pixels produce a description of the individual shapes. So the description of the individual shapes does involve basic grouping, similarity, and other principles common to the visual system. However, this level of description at the pixel level is completely detached from the higher levels in our stimuli. There is no applicable link between the pixel level and the dependencies that operate at the level of shape combinations regardless of which dependencies were used for generating the description of the individual shapes. The second level of control we used was to make sure that, at the level of the shape sets, there was no obvious way to use any of the preexisting pixel-level dependencies. The shapes were arranged in a standard grid with no variation in intershape distance, and each participant received an individual assignment of the shapes so that none of the shape base pairs could be more salient than others because of some accidental arrangements. Because our study focuses on the level of shape combinations, we argue that the current setup allows us to conduct this investigation in the most controlled way.

The main proposition of the current study is that the type of learning we measure in our paradigm is purely statistical in nature and, therefore, applicable on all levels of the visual system, from very basic contrast variations to high-level co-occurrences of objects and events. This is because the learning is operating on internal representations; therefore, it is irrelevant whether these representations represent pixels, blobs, or events. The traditional research on scene perception is heavily related to semantic meanings, associations between objects in the scene, and, in general, high-level cognitive functions. We use a much lower level operational concept of scene perception that is strongly influenced by infant research. For us, a scene is a complex set of elements in which many elements are highly visible, but most of them do not have categorical meaning or known relations to each other. Thus, infants could see blobs, edges, color patches, or changes in motion, and, based on these already available descriptions, they may develop a more precise understanding of which of these bits of information go together. This notion of scene perception is broad enough to encompass the visual problem for a 2-month-old as well as that for an adult, who naturally uses much more cognitive

(top-down) knowledge but, according to our view, the same principles. From this point of view, a scene is any visual input that has high complexity and unknown underlying structure. To validate this perspective, the next necessary step is to connect our results obtained with abstract shape-based inputs to the domain of low-level natural visual inputs and to demonstrate their applicability in real-world environments.

Related Work

Our model of constrained statistical learning is a direct descendant of Bayesian approaches (Knill, Kersten, & Yuille, 1996) and can be naturally formulated within a Bayesian framework (see the Appendix). Although there are studies that use the Bayesian approach for unsupervised learning (Dayan, Hinton, Neil, & Zemel, 1995; Frey, 1998), they must handle two basic problems confronting the full inference method of the pure Bayesian approach. First, they need to be able to scale up to real-sized problems without calculating all the possible likelihood functions of the task, which is impractical. This is essentially a restatement of the "curse of dimensionality" problem. Second, they need to avoid ad hoc narrowly defined goals as a cost function to their ideal observer model, thereby essentially custom fitting the learning model by using a complex model of the entire visual system (in fact, the entire brain). One approach to circumventing both shortcomings of the full inference method combines the Bayesian formulation of an ideal observer with the imitation of natural selection by a Bayesian formalization of evolutionary programming (Geisler & Diehl, 2002). Traditionally, developing an ideal observer model required a mapping of the entire space of all possible solutions to the problem. So learning the best features for visual recognition required a complete mapping of all situations in which that feature (and the other potential features) could be used and selecting the most effective set of the features. In contrast, natural selection does not map the entire space but changes the set of features in small steps, and with each step it improves the efficiency of the feature set incrementally. By combining the two methods, Geisler and Diehl (2002) posed constraints on the general process of finding the optimal feature set by introducing a general cost function through natural selection that is intuitively reasonable: maximum fitness of the species. The virtue of evolutionary programming in this framework is that it does not require a complete model of the entire problem space to search for a (locally) optimal solution. Conceptually, this approach is close to ours (cf. Mel & Fiser, 2000), because both of them explore the virtues of the constrained statistical approach to visual information processing, and both rely on the continuous feedback provided by the environment to improve the visual representation. The difference is that Geisler and Diehl (2002) provide a formal framework based on Bayesian statistical decision theory, whereas the current work explores empirically a set of specific constraints or shortcuts that seem to dominate visual feature development and thus should be incorporated in the formal framework.

Our findings are also related to the question of how humans form and store memory traces and retrieve those traces for recall, recognition, and decision making. These questions span a huge literature (Raaijmakers & Shiffrin, 1992; Ratcliff & McKoon, 2000; Squire, 1992; Yonelinas, 2002). As in these studies, the current work tests familiarity of previously seen patterns. However, in contrast to most of these studies of human memory, we focus on how parts of previously seen scenes are remembered. Although there were previous studies exploring the question of how humans remember parts of scenes or objects (e.g., Biederman, 1987), our work is, to our knowledge, the first to allow precise control over the statistics of the test scenes. In addition, because of our design (complete balance of statistical and semantic significance of the elements, high repetitiveness of the elements, uniform layout), our results cannot be explained by previous (prefamiliarization) experience and memories, episodic memory, or saliency. Therefore, our results can be more readily linked to the formation of general internal representations than to the ability to store particular memory traces. In other words, our work is more closely related to the particular problem of how episodic traces get transformed into knowledge representations and how relevant structures (i.e., features) get selected from visual experience and represented in memory in an efficient (i.e., nonredundant) manner.

These questions are also related to studies that focus on higher level aspects of feature learning, including the task dependency of feature learning for categorization (Goldstone, 2000; Schyns, Goldstone, & Thibaut, 1998) and aspects of causal learning (Gopnik et al., 2004; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Our approach shares the emphasis on the importance of constraints and the general probabilistic framework with these other approaches, but it differs in focusing on the basic constraints on the development of lower level perceptual features and attempts to minimize the effect of task dependency. Therefore, our experiments investigate a complementary aspect of human learning to those studied in the lines of research just presented. It is clear that, had we given the participants different kinds of tasks, their performance would have been influenced substantially. For example, if we had asked the participants to look for pair relations in Experiment 1, no doubt the results would have been just the opposite from what we reported. However, our paradigm explored the basic mechanism of learning in which the effect of the task was as minimal as possible, acknowledging that laying on top of this basic mechanism are various modulations in performance resulting from particular tasks. Goldstone (2000) and Schyns et al. (1998) investigated the scope and strength of such modulating task effects, whereas we asked whether there was any basic mechanism that would provide the foundation of learning when little or no cognitive (top-down) factors are present. One could argue that such a situation never occurs in the real world because our mind is always controlled one way or the other by some task. However, if this were true, our experimental paradigm should have led to a null result because different cognitive states should promote different preferences for pairs and triplets, and this should cancel out any bias across a large number of participants. The fact that we obtained a clear significant result suggests that our assumption of a basic mechanism that exists independent of the influence of the task is correct.

Finally, a number of studies investigate more directly what the possible representations for object recognition are (Fukushima, 1980; Hummel & Biederman, 1992; Mel, 1997; Oram & Perrett, 1994; Riesenhuber & Poggio, 2000) and how to acquire them (Dayan et al., 1995; Poggio & Edelman, 1990; Wersing & Korner, 2003). The dominant issue under heavy debate on the representa-

tional level has been whether the visual system uses a structural-description type of representation with a limited-size three-dimensional "alphabet" based on nonaccidental features and their spatial relations (Biederman, 1987; Biederman & Gerhardstein, 1993) or a view-based snapshot representation (Bülthoff & Edelman, 1992; Bülthoff, Edelman, & Tarr, 1995). Accumulating evidence from human studies (Foster & Gilson, 2002; Vuilleumier, Henson, Driver, & Dolan, 2002) and awake monkey experiments (Logothetis & Sheinberg, 1996) has led to a consensus that the brain probably uses both (Tarr & Bülthoff, 1998), and our framework readily accommodates this view (Fiser & Aslin, 2001). In terms of learning visual features, all proposed learning methods of both basic and higher level features (Hoyer & Hyvarinen, 2002; Karklin & Lewicki, 2003) must rely on the detection of conditional probabilities between parts of the input image, and our results provide the critically necessary experimental evidence that humans automatically extract these statistical measures.

Conclusions

Conducting a number of psychophysical experiments, we found that humans use a particular implicit strategy, termed the *embeddedness constraint,* that dominates the formation of internal memories of fragments that compose unfamiliar hierarchically organized scenes. We argued that this constraint helps to avoid the curse of dimensionality in visual feature learning by limiting the input space and the complexity of the available structure for learning. In addition, it also suggests a natural way of partitioning the visual input into chunks based on the statistical coherence between constituents of the input. We propose that, with this and similar constraints, a statistical learning approach offers a unified bootstrapping framework for investigating the formation of internal representations in vision. Other specific constraints guiding the formation of internal representations at various levels of this bootstrapping process should be the topic of further research.

References

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

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science, 218,* 697–698.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research, 23,* 3327–3328.

Bellman, R. (1961). *Adaptive control processes: A guided tour.* Princeton, NJ: Princeton University Press.

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

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1162–1182.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2,* 32–48.

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

Bülthoff, H. H., & Edelman, S. Y. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science, 89,* 60–64.

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

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

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

Dayan, P., & Zemel, R. S. (1995). Competition and multiple cause models. *Neural Computation, 7,* 565–579.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science, 208,* 1181–1182.

Fahle, M., & Poggio, T. (Eds.). (2002). *Perceptual learning.* Cambridge, MA: MIT Press.

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

Field, D. J., Hayes, A., & Hess, R. F. (2000). The roles of polarity and symmetry in the perceptual grouping of contour fragments. *Spatial Vision, 13,* 51–66.

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature, 287,* 43–44.

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 Science, 99,* 15822–15826.

Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society B (London), 269,* 1939–1947.

Frey, B. J. (1998). *Graphical models for machine learning and digital communication.* Cambridge, MA: MIT Press.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202.

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 357,* 419–448.

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

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 86–112.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111,* 3–32.

Hoyer, P. O., & Hyvarinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research, 42,* 1593–1605.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99,* 480–517.

Jacobs, D. W. (2003). What makes viewpoint-invariant properties perceptually salient? *Journal of the Optical Society of America (A), 20,* 1304–1320.

Karklin, Y., & Lewicki, M. S. (2003). Learning higher order structures in natural images. *Network: Computation in Neural Systems, 14,* 483–499.

Karni, A., & Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Science, 88,* 4966–4970.

Knill, D. C., Kersten, D., & Yuille, A. L. (1996). Introduction: A Bayesian formulation of visual perception. In D. C. Knill & W. Richards (Eds.),

Perception as a Bayesian inference (pp. 1–21). New York: Cambridge University Press.

Kourtzi, Z., Tolias, A. S., Altmann, C. F., Augath, M., & Logothetis, N. K. (2003). Integration of local features into global shapes: Monkey and human fMRI studies. *Neuron, 37,* 333–346.

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

Marr, D. (1982). *Vision.* San Francisco: Freeman.

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.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381,* 607–609.

Oram, M. W., & Perrett, D. I. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks, 7,* 945–972.

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

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

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual-learning in visual hyperacuity. *Science, 256,* 1018–1021.

Pomerantz, J. R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization: Simplicity and likelihood principles. In K. R. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance. Volume II: Cognitive processes and performance* (pp. 36/31–36/46). New York: Wiley.

Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology, 43,* 205–234.

Ramachandran, V. S., & Braddick, O. (1973). Orientation-specific learning in stereopsis. *Perception, 2,* 371–376.

Ratcliff, R., & McKoon, G. (2000). Memory models. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 571–582). New York: Oxford University Press.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*(Suppl.), 1199–1204.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21,* 1–54.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453–489.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67,* 1–20.

Vuilleumier, P., Henson, R. N., Driver, J., & Dolan, R. J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience, 5,* 491–499.

Wersing, H., & Korner, E. (2003). Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation, 15,* 1559–1588.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517.

Appendix

A Bayesian Framework for Statistical Learning

A typical probabilistic approach to perceptual phenomena is to have a Bayesian formulation that specifies the information in images that allows the observer to perceive the environment (Knill et al., 1996). This formalism has been used to define the available information for an ideal observer to solve a particular visual task and seems to be suitable for analyzing our observational learning paradigm. Briefly, in a classical Bayesian formulation, each scene S has a prior probability function $p(S)$, which represents the statistical dependencies between scene features. The observer's goal is to identify S based on a perceived image I. The S scene gives rise to the I image according to an image formation model π, with some additional noise N generated during the image formation: $I = \pi(S) + N$. The a posteriori conditional distribution $p(S|I)$ specifies the information that the image I provides about the scene S, and thus this is the quantity the observer seeks to obtain for successfully recognizing the input. According to Bayes's rule, this can be done by

$$p(S|I) = \frac{p(I|S)*p(S)}{p(I)}$$

where $p(I)$ is just a normalizing factor that represents the probability that image I occurs. Thus, in essence, the two terms of the numerator represent the factors that determine $p(S|I)$. The first, $p(I|S)$, is called the likelihood function, and it specifies the relative probabilities that different images appear for a given S. This term incorporates the effects of the image formation model π and the noise N. The second term, $p(S)$, is the prior distribution of scene configurations. This term collects effects of scene properties, the structure of the scenes, such as rigidity or smoothness, or the co-occurrences of elements.

In a real-world situation, both terms are important, although the relative importance might vary from case to case, and different formulations of the problem of perception emphasize one term or the other as being more important. In the classical signal-processing framework, in which the task is to identify the noise-corrupted image, the likelihood function is the critical term, and the $p(S)$ prior is typically substituted with a constant function assuming no additional relevant information attached to it. In contrast, the observational learning paradigm posits that noise and ambiguous image formation are a lesser problem, and the real challenge is in understanding the scene structure that is represented by $p(S)$.

Our observational learning task demonstrates in a nutshell how acquiring the conditional probabilities between elements interacts with this Bayesian framework. Because the displays in our task are composed of multiple shapes, identifying $p(S)$ means knowing the probability of appearance for each shape (12) at each location (5 * 5), which equals 300 individual entries for the distribution $p(S)$. Because of the combinatorial nature of the scenes with 12 possible elements, 25 possible positions, and 6 shapes in each scene, it also means that the learning method needs to consider $N = (12!/(6! * 6!)) * (25!/19!)$ possible individual scenes, which totals more than $1.2 * 10^{11}$ scenes in this small toy problem. One can assume identical probabilities for each of the shape entries at each position, but this does not reflect the true structure of the scene and leads to an inferior encoding and representation of the underlying structure. When the base-pair structures are noticed because of the conditional probabilities between the elements

of the base pairs, only six base pairs need to be encoded, and they can appear only in 20 or 16 positions (depending on the type of base pair) giving only 112 possible entries and approximately 6,000 possible scenes. The reason for this large reduction in entries and possible scenes is that now only a restricted combination of elements can make up any of the scenes because of the interaction between the higher order features (how pairs can be put next to each other) and the even smaller number of permissible scenes. Thus, by identifying the higher order features through the conditional probabilities, the complexity of the problem of perceiving and recognizing the scenes is substantially simplified.

Received January 26, 2005
Revision received June 22, 2005
Accepted June 23, 2005 ■

New Editors Appointed, 2007–2012

The Publications and Communications (P&C) Board of the American Psychological Association announces the appointment of three new editors for 6-year terms beginning in 2007. As of January 1, 2006, manuscripts should be directed as follows:

- *Journal of Experimental Psychology: Learning, Memory, and Cognition* (www.apa.org/journals/xlm.html), **Randi C. Martin, PhD,** Department of Psychology, MS-25, Rice University, P.O. Box 1892, Houston, TX 77251.

- *Professional Psychology: Research and Practice* (www.apa.org/journals/pro.html), **Michael C. Roberts, PhD,** 2009 Dole Human Development Center, Clinical Child Psychology Program, Department of Applied Behavioral Science, Department of Psychology, 1000 Sunnyside Avenue, The University of Kansas, Lawrence, KS 66045.

- *Psychology, Public Policy, and Law* (www.apa.org/journals/law.html), **Steven Penrod, PhD,** John Jay College of Criminal Justice, 445 West 59th Street N2131, New York, NY 10019-1199.

Electronic manuscript submission. As of January 1, 2006, manuscripts should be submitted electronically through the journal's Manuscript Submission Portal (see the Web site listed above with each journal title).

Manuscript submission patterns make the precise date of completion of the 2006 volumes uncertain. Current editors, Michael E. J. Masson, PhD, Mary Beth Kenkel, PhD, and Jane Goodman-Delahunty, PhD, JD, respectively, will receive and consider manuscripts through December 31, 2005. Should 2006 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2007 volumes.

In addition, the P&C Board announces the appointment of **Thomas E. Joiner, PhD** (Department of Psychology, Florida State University, One University Way, Tallahassee, FL 32306-1270), as editor of the *Clinician's Research Digest* newsletter for 2007–2012.