The effect of distributional information on feature learning

Joseph L. Austerweil (Joseph.Austerweil@gmail.com)
Thomas L. Griffiths (Tom_Griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650 USA

Abstract

A fundamental problem solved by the human mind is the formation of basic units to represent observed objects that support future decisions. We present an ideal observer model that infers features to represent the raw sensory data of a given set of objects. Based on our rational analysis of feature representation, we predict that the distribution of the parts that compose objects should affect the features people use to infer objects. We confirm this prediction in a behavioral experiment, suggesting that distributional information is one of the factors that determines how people identify the features of objects.

Keywords: representational change, features, rational analysis, Bayesian modeling

Introduction

> I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour... And yet even though such droll calculation were possible and implied, say, for the house 120, the trees 90, the sky 117 – I should at least have *this* arrangement and division of the total, and not, say, 127 and 100 and 100; or 150 and 117.
>
> Wertheimer (1938, p. 71)

A fundamental problem faced by any learner is the formation of the basic units that represent observed stimuli and support generalizations from a set of primitives. Wertheimer (1938) describes a visual form of the problem: how does the perceptual system form larger representations of observed objects from the information given by varying primitive units? Although the investigation of Gestalt principles has led to a fruitful body of research, there currently does not exist a formal computational account of why people form representations for novel objects. In this paper, we present a formal model of how feature representations should be inferred from a set of observed objects and demonstrate that people use statistical cues to infer the same features to represent novel objects that our ideal observer model would infer.

There are many factors that influence the features people infer to represent objects, like the changes of concavity of its contour (Hoffman & Richards, 1985), the usefulness for explaining categorization of objects (Schyns & Murphy, 1994; Pevtzow & Goldstone, 1994), background knowledge of the function of objects (Lin & Murphy, 1997), and prior knowledge of what types of features have been useful in the past (e.g., Gestalt principles, Palmer, 1977). However, we will focus on one particular factor: the distribution of features over objects. Intuitively, a feature representation is useful if knowing an unknown object has a feature gives you information as to which object it is. We propose an ideal observer model of feature inference that is sensitive to the distribution of parts over objects and demonstrate in a behavioral experiment that people infer features according to distributional cues as our model predicts.

A large body of previous research has demonstrated the powerful effect of statistical cues on human learning (Saffran, Aslin, & Newport, 1996; Aslin, Saffran, & Newport, 1998). Artificial language research has shown that human language learning faculties use the pattern of statistics of speech primitives to segment a continuous speech stream into words (Saffran et al., 1996; Aslin et al., 1998). We complement these results by performing a rational analysis of feature representation inference and demonstrating that people use statistical cues to infer feature representations for novel objects.

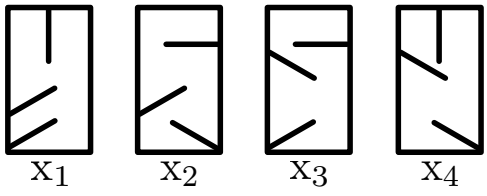
Our model is a nonparametric Bayesian model that allows for an unbounded amount of features to be expressed in the observed data. The model creates features to reproduce the objects it observes, but is penalized for each feature it produces. Thus, the model can infer the number of features necessary to represent the objects it observes. To the best of our knowledge, it is the only model of feature inference that infers the number of features from raw sensory data. Additionally, it has been shown to use distributional and categorization cues as people do (Austerweil & Griffiths, 2009).

This model makes a prediction based on how distributional information should affect the features people infer, which we now test in a behavioral experiment. If the parts that comprise objects vary independently over objects, then an observer should infer the parts as features. On the other hand, if the parts that compose objects covary over objects, an observer should infer the objects themselves as features.

The plan of the paper is as follows. In the first section we discuss previous empirical and computational work on human feature inference. Next, we present our ideal observer model and its predictions based on distributional cues. Third, we demonstrate people use the distributional cue as our model predicts in a behavioral experiment. Finally, we discuss the implications of our work for the nature of human concepts and future directions for research.

Inferring features

Perceptual and conceptual cognitive psychologists have been investigating the features people use to represent the world and both have been interested in how the features are created and change, for reviews of results from both fields see Goldstone (2003) and Schyns, Goldstone, and Thibaut (1998). To distinguish between the parts that exist in the objects and the features people use to represent observed objects, we will use "part" or "primitive" to refer to the aspect of the object and "feature" to refer to the representation of that object.



x_1	1	1	1	0	0	0
x_2	0	1	0	1	0	1
x_3	0	0	1	1	1	0
x_4	1	0	0	0	1	1

Figure 1: The four objects used in Shiffrin and Lightfoot (1997) and their feature ownership matrix.

Empirical Studies

Previous research has demonstrated two major influences of human perceptual feature learning: categorization and distributional information. In general, people infer features to represent objects that are useful for categorizing (the *functionality principle* of Schyns & Murphy, 1994). For example, Pevtzow and Goldstone (1994) demonstrated that participants inferred the diagnostic features useful to categorize each object into its appropriate category. They trained two groups of participants to repeatedly categorize the same four objects into different category schemes (objects A and B in one category vs. A and C in one category). Participants who learned to categorize A and B together inferred the shared part of A and B as a feature and those who learned to categorize A and C together inferred the shared part of A and C as a feature.

In addition to categorization cues, Shiffrin and Lightfoot (1997) showed that the distribution of parts over objects can affect the feature representation participants infer. In their visual search experiment, participants searched for one of the objects shown in the top of Figure 1 in a scene where the other three objects were distractors. The objects were designed so that each object shares one line segment with every other object (and thus, two line segments must be known to discriminate between objects). At first, participants do not experience "popout," meaning that response time in a visual search is nearly independent of the number of distractors. Popout typically only occurs when the target and distractor differ in a single feature. Thus, the objects must differ by more than one feature in the participants' representations (most likely a conjunction of line segments). However, after about 20 days of training, participants in the experiment experience popout. Therefore their feature representation of the objects must have changed to be the objects themselves.

Computational Models

Schyns et al. (1998) identified the need for computational accounts that infer feature representations and are psychologically motivated. Two factors that are important for any psychologically plausible model of feature learning are (a) the number of features should not be specified *a priori* and (b) the features should be inferred from raw sensory data. Previous work by Ghahramani (1995) and Goldstone (2003) described models that infer feature representations from raw pixel values; however, both models require the number of features to be specified ahead of time. This is a serious issue because finding the appropriate number of features to use is a difficult part of the problem of inferring features. For example, it is clear what the best feature representations are of sizes four and six for the objects in Shiffrin and Lightfoot (1997), but which of these two representations is more appropriate? People are not given this information and thus a model of feature inference should not receive it either.

More recently, Orban, Fiser, Aslin, and Lengyel (2008) defined a Bayesian learning model of visual chunks that can be interpreted as a model of feature representation inference. By training participants on scenes where novel objects occur in groups, they showed people infer representations that capture correlations between the groups as their model predicts. Although their model does infer the dimensionality of its representations, it is given each scene pre-processed as a binary string of whether or not objects occur. It does not infer its features from raw sensory input.

A Rational Analysis of Feature Representation

We will outline, following Austerweil and Griffiths (2009), a rational analysis of inferring features from raw sensory data without pre-specifying a specific number of features. First, we formalize the problem as finding the best feature representation Z for a set of observed objects X. We define Z to be a feature ownership matrix, where $Z_{ik} = 1$ indicates that object i possesses feature k (as in the matrix in the bottom of Figure 1. The problem of inferring Z from X can be solved by applying Bayes' rule, with the posterior probability $P(Z|X)$ being given by

$$\hat{Z} = \arg\max_Z P(Z|X) = \arg\max_Z \frac{P(X|Z)P(Z)}{\sum_{Z'} P(X|Z')P(Z')} \quad (1)$$

where $P(Z)$ is the prior probability of the feature matrix, and $P(X|Z)$, the likelihood, indicates the probability of the observed data given these features. This splits the problem into two subproblems: finding a representation that conforms to our prior assumptions, $P(Z)$, and finding one that can reproduce the observed objects with high probability, $P(X|Z)$.

As a prior on feature ownership matrices, we chose a nonparametric Bayesian prior, the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2006). The IBP can be interpreted to be a probability distribution over feature ownership matrices with varying numbers of features. The probability of a particular feature ownership matrix under the IBP is:

$$P(Z) = \frac{\alpha^K}{\prod_{h=1}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K} \frac{(N-m_k)!(m_k-1)!}{N!} \quad (2)$$

where N is the number of objects, K_h is the number of features with history h (the history is the column of the feature interpreted as a binary number), K is the number of features, H_N is the N-th harmonic number, and m_k is the number of objects that have feature k. One sensible prior assumption is that we should favor feature representations with a smaller number of features. By choosing α such that $\frac{\alpha}{N} < 1$, the IBP captures this intuition because the $\left(\frac{\alpha}{N}\right)^K$ term decreases when the number of features of the representation, K, grows.

In addition to the prior probability on feature representations, we present two probability distributions to use for recreating the observed objects X given a feature ownership matrix Z depending on the representation of the raw pixels. If the raw pixels are real valued, then a linear-Gaussian model (Griffiths & Ghahramani, 2006) can be used and if the raw pixels are binary, then a noisy-OR model (Wood & Griffiths, 2006) can be used. Using the noisy-OR model, Austerweil and Griffiths (2009) demonstrated that the model uses distributional and categorization information to infer representations as people do in both Pevtzow and Goldstone (1994) and Shiffrin and Lightfoot (1997).

One prediction the model makes is that when the parts weakly covary over objects (like those in Shiffrin & Lightfoot, 1997), objects should be inferred as features, but when the parts occur independently over objects, the parts should be inferred as features. It has not been shown yet that people use distributional information as the latter prediction suggests. Additionally, the rational analysis predicts people should infer objects as features even after observing the set of objects only a small number of times. To test the predictions of our model, Experiment 1 investigates how people infer feature representations after observing sixteen novel objects whose parts either weakly covary or are independent.

Testing the predictions: Martian Inscriptions

The goal of the experiment was to test the prediction of our rational analysis: when primitives are *correlated* over observed objects, people infer the objects as features, and when primitives are *independent* over observed objects, people infer the primitives as features. To investigate this prediction, we show participants a group of objects and look at how willing they are to call a new object that is a combination of three primitives a member of the previous group. According to our model, participants in the *independent* group should generalize to this new object (as they should infer the primitives), but participants in the *correlated* group should not (as they should infer the objects they observe as features and these cannot be combined to form the new object).

There were three between-subjects factors each with two levels: *distribution type* (*correlated* or *independent*), *training set* (1 or 2, which represents which of the primitives were correlated), and *test order* (1 or 2, which represents which of the two random orderings of the test stimuli were shown to participants). There was also one within-subjects factor with three levels: *test type*, indicating whether the test objects were previously *seen*, previously *unseen*, or made of *shuffled parts*.

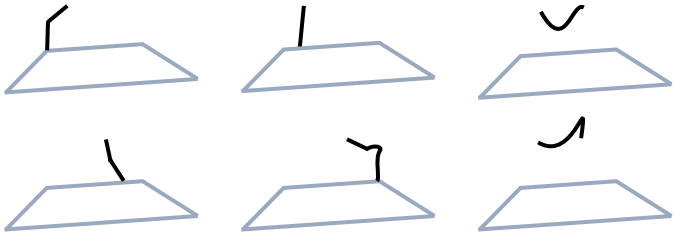


Figure 2: The six primitives used to create objects. The bias, which was in all objects, is shown in gray for reference, and the primitives are in black. Any combination of three features forms a connected object when combined with the bias.

Methods

Participants A total of 56 undergraduates from the University of California, Berkeley participated in exchange for course credit. There were 28 participants in each of the *correlated* and *independent* conditions with *training set* and *test order* counterbalanced.

Stimuli Figure 2 shows the images of the primitives and bias used to create the objects shown to participants. The objects were created by combining three primitives with the bias and were binary images. The primitives were designed such that any combination of three with the bias was connected, and so that people would have minimal prior knowledge (e.g., from Gestalt principles).

There were two *distribution types*: *correlated*, where primitives covary imperfectly over objects, and *independent*, where primitives were combined independently over objects. There were twenty possible objects, corresponding to all possible ways of choosing three features from a set of six. The *correlated* sets of objects were created to have the same correlation over primitives as Shiffrin and Lightfoot (1997) (see Figure 1). Two correlated sets were created using disjoint combinations of primitives, so that different objects appeared in each set. Each set consisted of four copies of four objects each with its own random added noise. The *independent* sets consisted of sixteen of the twenty possible objects. Again, two *independent* sets were created, with the four objects missing from each set corresponding to the four objects contained in one of the *correlated* sets. This method of generating stimuli guaranteed that each *correlated* set had a corresponding *independent* set in which each primitive appeared with the same frequency, allowing us to control for familiarity. Finally, noise was added to all of the images by flipping each pixel in the image with probability $\frac{1}{75}$.

Each participant was shown a training set – *correlated* or *independent* – with the specific set of objects depending on which *training set* condition they were in. Figure 3 shows the images in one *independent* and one *correlated* condition. Participants viewed their training set by exploring the objects

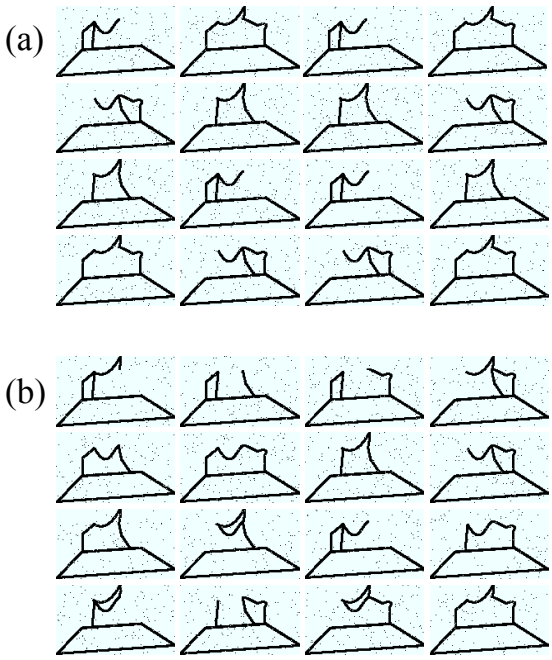


Figure 3: One of the *correlated* and one of the *independent* training sets given to the participants. (a) One of the two *correlated* training sets. (b) One of the two *independent* training sets. These two sets share four objects.

printed on cards, as described in more detail below. The same test set was given to all participants in one of two random orders. There were twelve objects in the test set, as shown in Figure 4. The twelve objects fell into three *test types*: four objects seen by the participant already (*seen*), four objects the participant had not seen already that were composed of the same primitives (*unseen*), and four objects created by combining primitives inconsistent with the statistical information from both training sets (*shuffled parts*). As a consequence of the way the stimuli were constructed, the *seen* and *unseen* test objects corresponded to one of the two *correlated* sets – which objects participants had seen or not was determined by the *training set* condition. This allowed us to control for the possibility that one set of objects was naturally more appealing than the other. The *shuffled parts* tests were created by first taking the image formed by joining all six parts and segmenting it into six different parts. The *shuffled parts* images used in the tests were four objects formed by a combination of three of the six shuffled parts. This was done so that the four shuffled images would have the same gross properties as the other test images.

The stimuli and test sets were carefully constructed to ensure that: (1) the variance at each pixel was equal for all training sets, (2) the features that were used in constructing the correlated set were counterbalanced, and (3) the average similarity (in terms of pixel overlap) between any training set and any test set was equal.

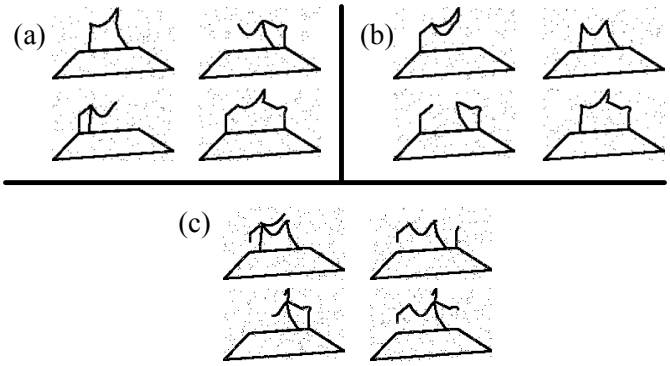


Figure 4: The three sets of test images. (a) *seen* for training set 1 (shown in Figure 3) and *unseen* for training set 2. (b) *unseen* for training set 1 (shown in Figure 3) and *seen* for training set 2. (c) *shuffled parts* for all conditions.

Procedure Participants were given the sixteen images on business cards randomly shuffled in front of them appropriate to their conditions and given the following cover story:

Recently a Mars rover found a cave with a collection of different images on its walls. A team of scientists believe the images could have been left by an alien civilization. The scientists are hoping to understand the images so they can find out about the civilization.

They were asked to alert the experimenter after "investigating the images" by "laying all the cards out on the table and organizing them in any way you think might help you learn about the images" and told that "no longer than 5-10 minutes is necessary." After they finished investigating the images, they were given the following test instructions:

It looks like there are many more images on the cave wall that the rover has not yet had a chance to record. If the rover explored the cave wall further, which images do you think it would be likely to see?

Your task is to rate how likely you believe it is that the rover sees each image as it explores further through the cave.

In the booklet in front of you are twelve images, each on its own page. After you are finished rating each image, turn the page to the next image. Once you have turned to the next image, please DO NOT TURN BACK to any previous images.

To minimize memory effects, the images from the training set were not taken away from the participants. Each image was shown on a single page and participants were asked to generalize to the test set ("rate from 0-10 how likely you believe the rover is to see this image on another part of the cave wall").

Results

Figure 5 shows mean responses and model predictions. Participant responses were grouped into the three *test types* (*seen*, *unseen*, and *shuffled parts*) and then averaged. Model predictions were calculated from the probability of the new images given the images from either the independent or correlated conditions, and averaged in the same way. The model

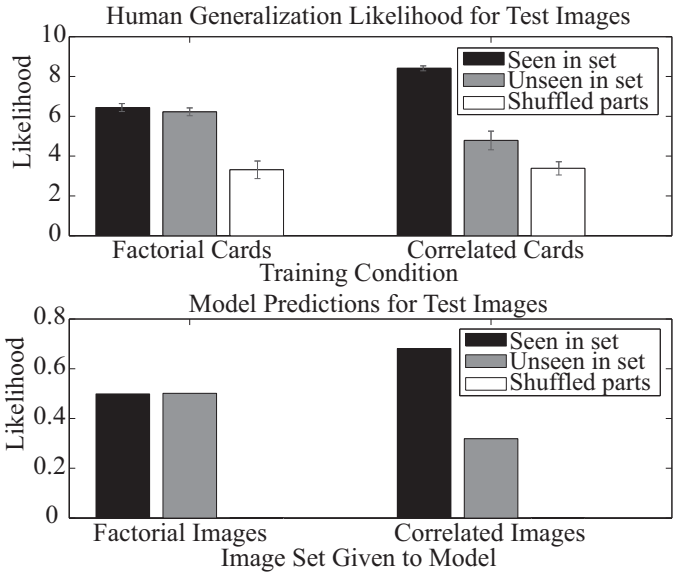


Figure 5: Experiment results. The upper panel shows mean ratings participants for test items as a function of training condition. Error bars show one standard error. On bottom, model predictions for the same test images given images from either the independent or correlated conditions.

predictions were computed by approximating the full posterior predictive distribution with the probability of the new images using the most likely features as determined by a Markov chain Monte Carlo simulation (see Austerweil & Griffiths, 2009 for details). Since there was a large difference in the probabilities of different types of test images, we use a monotonic but non-linear transformation to produce the values shown in the plot, raising the probabilities to the power of 0.0005 and renormalizing. Qualitatively, the model and people show the same pattern of responses on all test items.

A four-way ANOVA revealed a main effect of *test type* ($F(2,52) = 61.01, p < 0.001$), an interaction between *test type* and *distribution type* ($F(2,52) = 10.57, p < 0.001$), and no other significant main effects or two-way interactions (all $F < 1$). There was a three-way interaction of *test type*, *test order*, and *distribution type* ($F(2,52) = 19.11, p < 0.05$). However, the effect is irrelevant to the question of whether people use distributional information as it is caused by participants in the first *test order*, *independent* condition rating the seen images higher than those in the second *test order*, *independent* condition. Since there were no major effects of *training set* or *test order*, we collapsed over these conditions in the subsequent pre-planned analyses.

Confirming our hypothesis, participants in the *independent* condition are more likely to generalize to the unseen images than those in the *correlated* condition ($t(54) = 3.05, p < 0.005$). There was no difference between the *seen* and *unseen* image ratings for the participants in the *independent* condition ($t(54) = 0.27, p > 0.05$); however, there was for those in the *correlated* condition ($t(54) = 8.74, p < 0.001$). Participants in the *correlated* condition were more likely to generalize to the *seen* images than those in the *independent* condition ($t(54) = 2.97, p < 0.005$). Participants in both training conditions are more likely to generalize to the *seen* images than the *shuffled parts* images ($t(54) = 10.07, p < 0.001$ and $t(54) = 4.63, p < 0.001$ respectively). There was no difference between participiants in training conditions on the *shuffled parts* images ($t(54) = -0.12, p > 0.05$). Finally, participants in both the *correlated* and *independent* conditions are more likely to generalize to the *unseen* images than the *shuffled parts* images ($t(54) = 2.89, p < 0.01$ and $t(54) = 5.31, p < 0.001$, respectively).

Discussion

The main results of our experiment confirm the predictions of our model: participants in the *independent* condition do not differentiate between the *seen* and *unseen* images; however, participants in the *correlated* condition do. Additionally, participants in the *independent* condition are more likely to generalize to the *unseen* objects than those in the *correlated* condition. Since participants in the *correlated* condition should expect fewer objects under the feature representation predicted by our model (just the four objects they observed), it is sensible that they rate the *seen* objects higher than the *independent* group. Finally, both groups rate the *shuffled parts* images lower than the *seen* and *unseen* images.

Participants in the *independent* group generalized to the *unseen* objects, while those in the *correlated* group did not. Neither group generalized to the *shuffled parts* objects and there is no significant difference between the groups on the *shuffled parts*. Our results cannot be explained by participants in the *independent* group just expecting more variance in test objects than those in the *correlated* group. First, as noted above, the variance at each pixel was equal across training sets. Second, if participants in the *independent* group simply expect more variance, this should predict that they would be more willing to generalize to the *shuffled parts* as well as the *unseen* objects, which was not the case. The pattern of judgments on the different test items made by participants in the two groups also cannot be explained by a simple categorization model with the pixels as features because it would not distinguish between the types of test items due to the way the training and test sets were constructed: the similarity (in pixel overlap) was equal between all training and test sets. Thus, our results suggest that participants infer features appropriate to the distributional cues between parts in the objects they observe.

One might argue that participants in the *correlated* condition still differentiate between the *unseen* and *shuffled parts* images and that this in some way invalidates the predictions of the model; however, most of the images in the *shuffled parts* set are poorly formed according to Gestalt principles and our model does not take into account these effects. In a follow-up experiment, we are creating a new *shuffled parts* set that does not violate our prior notions of what a good object looks like. Additionally, one might argue that these results are

due to some aspect of the particular primitives we correlated together (e.g., they form some pre-existing salient object), but since there was no effect of *training set*, we demonstrate this was not the case. Since participants in both the *correlated* and *independent* conditions observe the parts the same number of times throughout the object set, they must be sensitive to the covariation of parts in objects and not just the overall occurrence of the parts themselves.

General Discussion and Conclusions

We have demonstrated that the statistics of how parts vary over objects affects the features inferred by participants. Based on our ideal observer model, we predicted participants should infer the parts of novel objects as features when they occur independently over objects and the objects themselves as features when they covary. Participants who observe a set of only 16 objects whose parts covary do not believe an unseen valid combination of parts is a member of the original set; however, those observe a set of 16 objects whose parts occur independently do believe the same unseen valid combination of parts is a member of the original set. Thus, people use statistical cues to infer features that represent objects, which influence later decisions about the objects.

Is this effect something unique to visual perception, or does it reflect a general cognitive ability to appropriately extract parts or wholes of objects as features? The previous work demonstrating the importance of statistical cues for inferring words and actions suggests it is a general cognitive capacity. To test this, we hope to run a follow-up experiment using the same paradigm in a conceptual domain.

Our analysis provides a principled computational framework to investigate this problem, identifies key factors influencing the learning of feature representations, demonstrates people use these factors in the same way as an ideal observer, and predicted a new empirical result in how people infer feature representations. In addition to furthering our knowledge of human feature learning, our results are important to machine learning because the problem of representing the world in a useful way is shared between machine learning and cognitive psychology. Finally, our results are first steps towards a larger goal. We hope to extend our computational model to capture the effects of categorization, causality, relations, and prior knowledge and how people infer features.

Acknowledgments. We thank Rob Goldstone, Stephen Palmer, Karen Schloss, Tania Lombrozo, Charles Kemp, Noah Goodman, and Eleanor Rosch for insightful discussions, Amy Perfors and three anonymous reviewers for comments, and Brian Tang and David Belford for help with experiment construction, running participants, and data analysis. This work was supported by grant FA9550-07-1-0351 from the Air Force Office of Scientific Research.

References

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

Austerweil, J. L., & Griffiths, T. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, Y. Bengio, D. Schuurmans, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems* (Vol. 21). Cambridge, MA: MIT Press.

Ghahramani, Z. (1995). Factorial learning and the EM algorithm. In *Advances in Neural Information Processing Systems* (Vol. 7, p. 617-624). Cambridge, MA: MIT Press.

Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In *Perceptual Organization in Vision: Behavioral and Neural Perspectives* (p. 233-278). Mahwah, NJ: Lawerence Erlbaum Associates.

Griffiths, T. L., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18). Cambridge, MA: MIT Press.

Hoffman, D. D., & Richards, W. A. (1985). Parts in recognition. *Cognition*, *18*, 65-96.

Lin, E. L., & Murphy, G. L. (1997). Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 1153-1169.

Orban, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745-2750.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*, 441-474.

Pevtzow, R., & Goldstone, R. L. (1994). Categorization and the parsing of objects. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (p. 712-722). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Schyns, P. G., & Murphy, G. (1994). The ontogeny of part representation in object concepts. In *The Psychology of Learning and Motivation* (Vol. 31, p. 305-354). San Diego: Academic Press.

Shiffrin, R. M., & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. In *The Psychology of Learning and Motivation* (Vol. 36, p. 45-82). San Diego: Academic Press.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. Ellis (Ed.), *A source book of Gestalt psychology* (p. 71-88). London: Routledge and Kegan Paul.

Wood, F., & Griffiths, T. L. (2006). Particle filtering for nonparametric Bayesian matrix factorization. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems* (Vol. 18). Cambridge, MA: MIT Press.