Grammar-based object representations in a scene parsing task

Virginia Savova, Frank Jäkel, Joshua B. Tenenbaum
[savova, fjaekel, jbt]@mit.edu
Brain and Cognitive Science, MIT
Cambridge, MA 02139

Abstract

This paper addresses the nature of visual representations associated with complex structured objects, and the role of these representations in perceptual organization. We use a novel experimental paradigm to probe subjects' intuitions about parsing a scene consisting of overlapping two-dimensional objects. The objects are generated from an abstract 2-dimensional image grammar, which specifies the set of possible configurations of object parts. We show that participants' performance on the task depends on prior experience with the object class, and is based on structural cues. This indicates that structural representations exerted a top-down influence on parsing. To address the question of representation type, we used a computational model of object matching in conjunction with various probabilistic representational models. Our simulations indicate that grammar-based representations derived from the original grammars are superior to more restrictive exemplar-based representations in explaining human performance on this task, as well as to more inclusive, over-generalizing grammar-based representations.

Keywords: visual representations, computational modeling, object categories

Introduction

Conscious visual perception is strikingly removed from the patterns of light which stimulate our eyes, or even the low-level visual features active in primary visual cortex. Rather than experiencing the world as a collection of edges, bars or patches of color, we perceive scenes, hierarchically composed of objects, and parts. Compositionality is one of the defining characteristics of visual representations, a fact evidenced by our ability to predict the location of an occluded part, identify an object despite change of relative part position, or imagine a novel object as a combination of existing object parts. The nature of these compositional representations raises a series of questions regarding their origin and limitations.

A cursory examination of just a few common visual object categories reveals various degrees of structural complexity. On one end of the spectrum, we find simple categories, such as ball, or plank, which have a degenerate compositional structure, consisting of a single part. In a probabilistic framework, individual objects within these categories may well be thought of as samples from a probability distribution over a feature space of homogeneous, smoothly varying characteristics such as size, shape, color. The probability distribution over the feature space forms the representation of the category. The middle of the spectrum is occupied by those object categories which are typically the target of computer vision algorithms and applications, such as faces, cars, bicycles or tools. These objects exhibit compositional structure. The building blocks are relatively autonomous parts which enter into specific spatial relationships with each other and form equivalence classes. The representation of individual parts is analogous the the representation of simple objects. However, the representation of the objects as a whole requires an additional specification of part-part, or part-whole spatial relationships, which also yields itself to a probabilistic interpretation. We will refer to these representations as part-based. Finally, on the rarely examined far end of the spectrum we may discover a variety of objects whose compositional structure cannot be satisfactorily captured by part-to-part relationships, probabilistic or otherwise. Categories of this type include houses, churches, circuit boards, molecules, and various life-forms, most notably trees, as well as many plants and sea creatures. Unlike mid-spectrum categories, these objects have a variable number of subparts, and their parts may be hierarchically composed of other subparts of unbounded depth, which complicates the specification of template-like part-part and part-whole relationships. Elsewhere, we have argued that such object classes are best represented as the extension sets of grammar-based generative models (Savova & Tenenbaum, 2008).

If we assume that the visual properties of these categories are faithfully represented, a more powerful representational system, grammar-based or grammar-equivalent, would be required. However, the mere existence of structural complexity in the distal stimulus does not provide direct evidence for representational complexity in the visual system. It is possible that the representation of these objects is impoverished as compared to their actual structural characteristics. Alternatively, complex structured objects may be the exception, rather than the rule, and we might be limited in our ability to acquire novel categories of this type naturally or effortlessly. In so far as full structural representations do exist, they might be of a post-visual, purely conceptual nature, and not indicative of the representations engaged during normal visual processing. For example, representing the full spatial structure of houses is not required for recognition or detection. Under normal circumstances, a few diagnostic features might do. Therefore, it is essential to explore the nature of visual representations in the context of a visual task which directly engages the entire object representation.

The idea that objects and scenes are represented in terms of structural descriptions involving parts and spatial relations is not new (Marr & Nishihara, 1978) (Palmer, 1999), and neither is the idea that the formation of this representation is guided by the statistical properties of the input. Structural models are a classic idea in the field of cognitive psychology. For example, the recognition-by-components theory sought to describe objects as 3D simple shapes composed with binary spatial relations (Bierderman, 1987). Structural template models such as constellation models (Fergus, Perona, & Zisserman, 2003) have recently become popular in computer vision, because they allow some deformation by encoding a distribution on the relative position of parts. However, few researchers have explicitly addressed the question from the point of view of a generative object model, in particular the kind of model capable of generating complex structural descriptions involving recursion and unbounded derivational depth. Grammar-based models of vision have been anticipated in some classic works in computer vision and syntactic pattern recognition (Fu, 1974), but have received a new wave of attention relatively recently (Zhu, Chen, & Yuille, 2006; Wang et al., 2006) in a probabilistic setting.

While experiments have demonstrated the sensitivity of human adult and infant learners to object-level statistical information (Fiser & Aslin, 2001), the nature of the representations over which statistical learning operates is an open problem. Recently, there have been a few attempts to investigate the hypothesis that people learn about object categories by inferring a generative model in classification and similarity judgement tasks (Hegde, Bart, & Kersten, 2008).

It is not clear to what extent these tasks are solved on the basis of the full category representation, rather than some combination of salient diagnostic features or logical inference. To assess the nature of internal visual representations, we investigate to what extent such representations can be spontaneously acquired from visual information, and can subsequently be engaged. Usually, visual representations are investigated in recognition and categorization tasks. However, these tasks can often be solved without engaging full object representations, on the basis of discriminative features. More complicated tasks, such as parsing a scene are better suited for the study of complex representations.

Consider the town view in Figure 1. Local segmentation cues cannot account for the full set of parsing decisions that need to be made by the visual system in this context. We clearly perceive windows and doors as being on the houses, rather than floating in front of them; individual houses occluding one another, rather than standing on top of each other; smaller houses as being further away; the rooftops as the topmost parts of houses, rather than floors of the houses above them. To explore the shift in perception when inconsistency is detected, we can flip the image upside down. Notice that the grey rooftops are now parsed as above (and part of) the green-window house. This demonstration provides some evidence that our prior experience with houses informs the visual analysis of this image. Thus, scene parsing seems to tap into the structural representations of objects. Consequently, a controlled parsing task with novel complex structured objects may allow us to distinguish between exemplar-based representations and more powerful, grammar-based representations.

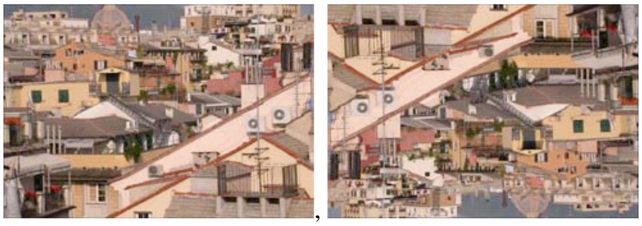

Figure 1: Segmentation decisions are informed by knowledge of object structure (Photograph by Markus Wiedemeier). Right: same photo upside-down.

Investigating complex structured objects in a parsing task

Models of representation

To investigate the nature of representations, we defined a set of generative models of increasing generality. We used a parsing task to test subjects' structural inference on stimuli generated from a grammar-based model, and assessed which of the models provides the best explanation for subject performance by matching samples from each model to the regions segmented by the subject.

The exemplar model is a trivial generative model which enumerates a set of objects. This model does not provide for explicit generalization, but generalization may be achieved indirectly in the matching procedure. The part model is a step above the exemplar model. It generates objects from a set of templates which provide part-part relationships by varying the size of parts independently. The grammar model uses a set of rules to recursively generate objects from parts. Unlike the part model, the grammar model is in principle capable of generating hierarchical object representations of differing or unbounded depth. Finally, the supergrammar model allows all configurations allowed by individual realizations of the grammar model.

In some sense, all of these models lie on a continuum. The exemplar model can be thought of as a degenerate part model, since it permits no independent variation among parts. The part model may be viewed as an impoverished grammar-based model, since a template is equivalent to a grammar which allows for only a finite set of fixed-depth derivations and associated hierarchical descriptions. The supergrammar model, on the other hand, is a generalization of the grammar model. Thus, each of the models generates a superset of the objects generated from preceding models.

A set of thirty grammars was randomly sampled from a metagrammar, which generated grammars by picking how many rules of each type (obligatory, optional or recursive) should be attached to each non-terminal, and the direction of expansion for each rule (up, down, left or right) from a uniform distribution. A typical grammar is shown in Table 1

The grammar model was used to generate visual objects in two steps. Each object was generated by a) generating a parse tree and b) varying the size of the vocabulary parts in-

Table 1: Typical grammar form

A	→	A	B	right	obligatory	
B	→	B	B	left	recursive	
B	→	B	B	right	recursive	
B	→	B	C	right	obligatory	
B	→	B	D	right	optional	
C	→	C	D	right	obligatory	
C	→	C	E	up	optional	
C	→	C	F	down	optional	

dependently according to a scaling factor drawn from a truncated Gaussian distribution with parameters $mean = 0.75$, $var = 0.25$, $min = 0.5$, $max = 1$. These parameters were chosen to favor smooth variation of each part within the category. Each symbol was expanded to a set of right-hand symbols by the following procedure:

1. Apply all obligatory rules for this symbol type.

2. Apply all optional rules with a fixed optional rule probability $p = 0.75$.

3. Apply all recursive rules with a fixed recursive rule probability $p = 0.5$.[1]

4. Enter all newly generated right hand symbols in the expansion queue.

Once all the rules were applied to a symbol in the expansion queue, the symbol was deleted from the queue. The parse tree was initialized with a unique start symbol.

Experiment

Three sets of objects were generated from each grammar: a stimulus set, containing between five and seven objects, an example set containing twelve objects, and a sample set of a hundred objects used for modeling purposes. Fifteen of the thirty grammars were subsequently chosen to span a continuum of two grammar complexity measures, one taking into account the number of rules, the other taking into account the presence of recursive rules. The stimulus set was used to generate a display of overlapping objects, where each object position was a sample from a horizontal and vertical beta distribution. The resulting scenes were artificial, simpler analogues to the natural scene in Figure 1.

The experiment included twelve participants total. Participants were told that they will be segmenting alien objects from NASA images which contain partially occluded objects. They were asked to outline where each object was approximately located. A pen tablet was used as a tool for drawing contours on the screen. Subjects were given time to practice with the tablet before starting the experiment. In a demo session, they were shown a cutout of a ballet scene with dancers partially occluding each other. They were shown the correct outline for each object. In a practice session, they were given three practice trials with practice stimuli drawn from simpler grammars which were not part of the stimuli in the experiment. In the test session, they were asked to complete fifteen trials, one per grammar. To ensure that participants treated the scenes as compositions of homogeneous objects, the objects were assigned a different nonce-word category name in the instruction preceding each trial.

For nine of the twelve subjects, each trial consisted of a study phase, which presented the subject with either a subset, or the full set of the examples generated from the grammar for this trial. Three, six, or twelve examples were used, depending on the group the subject was in. There were a total of three groups, ensuring that each display was presented with a different number of examples at least three times. After viewing the examples, the subject pressed a key to move to the parsing phase. Upon completion of a contour, they were given a chance to delete or accept it. Once a contour was accepted, it could no longer be deleted. The subject was free to move to the next trial whenever they considered the current parsing complete. Three additional participants were assigned to a fourth group, for which the same stimuli were presented without the study phase. These subjects were simply asked to outline the individual objects in each screen to the best of their abilities, without any additional information about the objects.[2]

Data analysis

The data was analysed against ground truth using the standard precision-recall analysis. First precision and recall were calculated for each object-contour pair. Precision was defined as

$$Precision = \frac{TP}{TP + FP}$$

where TP is the number of non-black pixels inside the contour which belong to the object, and FP is the number of pixels inside the contour that do not. Recall was defined as

$$Recall = \frac{TP}{TP + FN}$$

where FN is the number of non-black pixels outside the contour which belong to the object.

The harmonic mean of the two measures (the F-measure),

$$F = 2 * \frac{Precision * Recall}{Precision + Recall},$$

was then calculated for each pair, and contours were assigned to objects in a way that maximizes the F-measure for each object-contour assignment.

[1] this parameters were chosen to ensure that recursive and obligatory rules were applied a substantial number of times while keeping the probability of generating very large or infinite objects low. Any such objects were filtered out

[2] It was not possible to alternate between study phase trials and test-phase-only trials within subjects, for fear of introducing cross-trial transfer-learning effects

If the subject draws fewer contours than there are objects we can pretend the subject has drawn as many contours as there are objects but the contours do not contain any of the pixels of the objects, hence the F-score is always zero for these contours (Figure 2). Conversely, if the subject has drawn more contours more contours than there are objects we can say that the subject tried to circle an object that is not in the scene and therefore the F-score for this object is always zero.

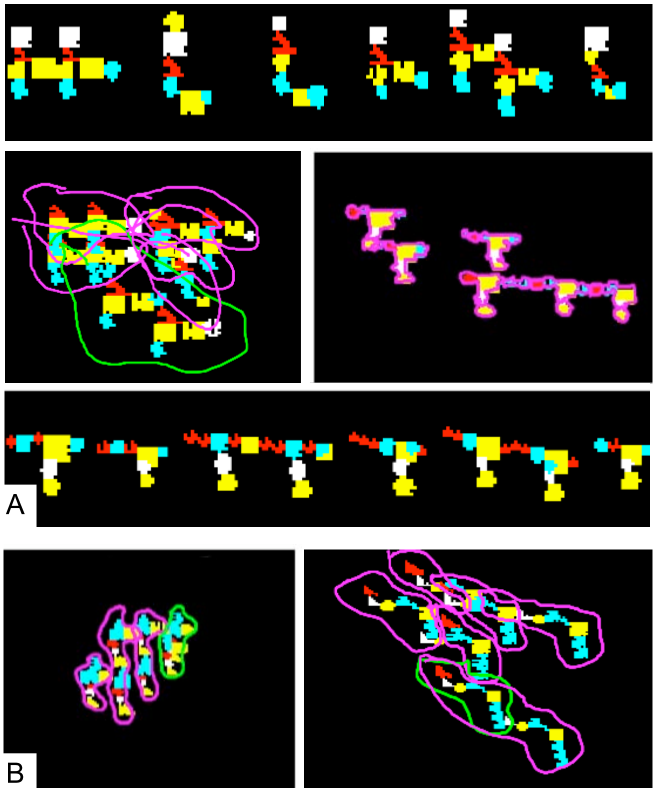

Figure 2: Actual subject parses. Panel A first and third row: examples from a grammar with relatively high variability (Grammar H) and examples from a grammar with low variability (Grammar L). Second row on the left: An actual screen with stimuli Grammar H showing a correct instance of generalization (in green). Second row on the right: An actual screen with stimuli from Grammar L. Panel B: A case of oversegmentation and undersegmentation (in green).

The mean F-measure per grammar for all subjects against example set size and grammar complexity, as measured by the number of recursive rules, and by the number of overall rules, are plotted in Figure 3. The example set size and the number or recursive rules were found to be significantly correlated with the F-score ($\rho = 0.32; \rho = -0.47, p < 0.002$). In addition, the correlation analysis revealed that the factors number of recursive rules, and number of overall rules were correlated. Therefore the exact source of the complexity effect cannot be conclusively determined.

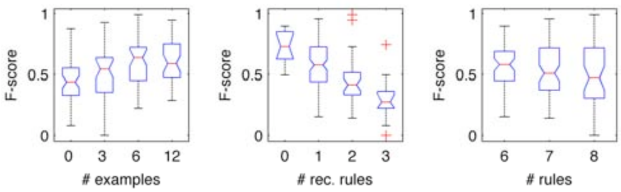

Figure 3: Box plots for example set size (left), number of recursive rules (center), and number of rules overall.

Simulation results

To address the question of what type of representations the subjects were using in generating contours, we sampled from the four generative models described earlier. The exemplar model for each subject and each trial "generated" exactly the object that person had seen during the study phase. For the subjects who saw no examples, the exemplar model was run with the full set of twelve examples, in order to assess the differential effects of human learning and the representativeness of the exemplars as a mini-sample from the grammar. The part model generated variants of the objects in the exemplar model by varying their parts according to the same scale factor distribution used in the grammar generative model. The grammar model was equivalent to the original model used to generate the stimuli. Finally, the supergrammar model was derived from all grammar models used in the experiment by concatenating all unique rules of their joint rule set and switching the status of obligatory rules to optional. Thus, the generative capacity of the supergrammar is a superset of the generative capacity of all used grammars. The same procedure was used for generating samples from the supergrammar as for each individual grammar. The supergrammar sample for each trial used the part-model specific for that trial.

One hundred samples were drawn from each of the models. Each sample was compared to the content of each contour drawn by each subject, and the mean likelihood of the samples from each model was computed according to the following matching model:

Let x be an image of a scene that shows all the non-black pixels contained in one contour produced by a subject and has zeroes everywhere else. We want to evaluate how well each model would be able to explain this object x. To this end we compare x against all stimuli s that could be generated by each model M_i. Hence, we calculate

$$p(x|M_i) = \sum_s p(x|s)p(s|M_i)$$

for all i by means of a Monte-Carlo approximation with the hundred samples of s from $p(s|M_i)$, generated earlier as described. Since x is an image of the size of the scene with some non-zero pixels at some position we need to take into account that the stimulus s could have appeared at all postions pos on the screen. Hence,

$$p(x|s) = \sum_{pos} p(x|s,pos)p(pos)$$

is obtained by sliding the stimulus s as a template over the whole display and assuming a flat prior.

Instead of modeling the exact relationship between the pixels in x and s we only model the counts of how many pixels agree or disagree with each other. Let N be the overall number of pixels in the segmented image of the scence x (including black pixels) and let K be the number of non-black pixels of the stimulus s. For a given position pos that we assume the stimulus to be at in the scene the number of hits h that the subject would have obtained is given by the number of non-black pixels in s that have been marked by the subject in x. Note that if the subject's contour includes all the pixels of s the number of hits h is exactly K. Since another object might overlap with s the actual color of a pixel in x might not be the same as in s and hence we do not require the pixels in x and s to have the same color. The number of misses is then $m = K - h$. Similarly, the number of false alarms f is given by the non-black pixels that were marked by the subject as belonging to s even though they do not overlap with the template. The number of correct rejections is then $c = N - K - f$. Assuming that there is a hidden parameter p that determines the probability of a hit and another parameter q that determines the probability of a false alarm, we model the probability of obtaining the subject's response x assuming that s is at position pos as:

$$p(x|s,pos,p,q) = \left(\frac{\Gamma(h+m+1)}{\Gamma(h+1)\Gamma(m+1)}p^h(1-p)^m\right) \cdot \left(\frac{\Gamma(f+c+1)}{\Gamma(f+1)\Gamma(c+1)}q^f(1-q)^c\right),$$

the product of two binomial distributions: one for the pixels that lie on the template and one for the pixels outside the template. Since the exact values of p and q are unknown we integrate them out using a prior. We first assume that p and q are independent of each other. The prior for p should be peaked at zero because if the subject assumes s to be the stimulus the contour that the subject draws should include all the pixels in s. The subject will make mistakes due to overlapping objects and sloppy drawing, however. A reasonable prior for p seems to be a beta distribution prefering values close to one but with a mean on the order of a high proportion $\gamma = 0.9$ of the pixels of the stimulus, i.e. $p(p) = Beta(p;\alpha,1)$ with $\alpha = 1/(1/\gamma - 1)$. As a prior for q we choose a beta distribution that prefers not to produce false alarms. Again the number of false alarms that are produced should on average be on the order of $1-\gamma$ times the number of pixels of the stimulus, i.e. $p(q) = Beta(1,\beta)$ with $\beta = \frac{N-K}{(1-\gamma)K} - 1$. Integrating out these priors we obtain

$$p(x|s,pos) = \alpha\frac{\Gamma(h+m+1)}{\Gamma(m+1)}\frac{\Gamma(h+\alpha)}{\Gamma(h+m+1+\alpha)} \cdot \beta\frac{\Gamma(f+c+1)}{\Gamma(c+1)}\frac{\Gamma(c+\beta)}{\Gamma(f+c+1+\beta)}$$

The results of the simulation are summarized by Figure 4. Each graph in the figure represents a histogram of the log odds of two models for all contours. The grammar model outperforms all other model for those subjects who have seen three examples or less, and continues to outperform both the exemplar model, and the supergrammar model for all other subjects. The comparison of the grammar model with the part model for subjects who have seen six examples or more reveals a slight advantage for the latter model which may be attributed to the fact that, as the number of examples increases, the part model becomes more similar to the grammar model in terms of the diversity of its sample. Since one-shot learning in visual tasks has been experimentally demonstrated (Fei-Fei, Fergus, & Perona, 2006), (Savova & Tenenbaum, 2008), the good performance of the grammar model with few examples is particularly important.

The limitations of the part model are illustrated in Figure 5. A typical contour drawn by a subject in the study condition included the long object in its entirety. The grammar model can find a much better fit for this contour than the part model, which is based on the template provided by exemplars. All subjects who were not provided with any exemplars segmented this object into several smaller objects.

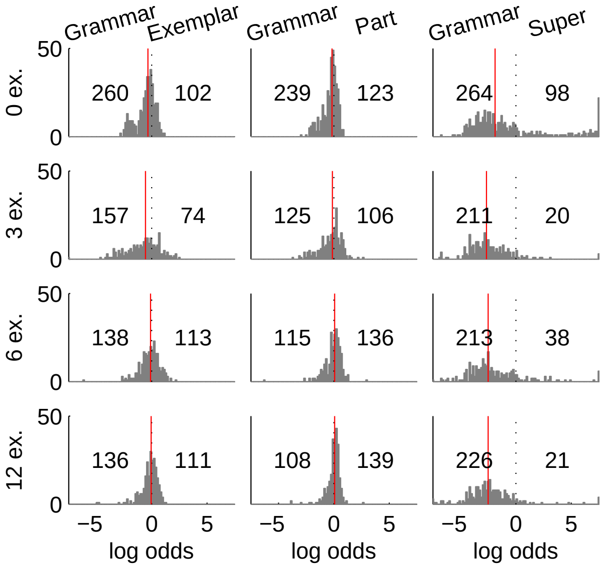

Figure 4: A comparison of the representational models. Log odds histograms. The red line is the median of the distribution. The dotted line is at zero. Negative values indicate higher log likelihood for the grammar model. Number of contours better fitted by each model are inside the plots.

Discussion

A basic finding of our experiment is that parsing depends on the amount of top down structural information available about the category, as evident by the effect of example set size on performance. If parsing was based on bottom-up cues alone, we would not expect such an effect.

The results of the model comparison illustrate the advantages and limitations along the continuum of increasing representational complexity. The exemplar model, which is a highly constrained sample of the grammar model, is easily overtaken by the part model which provides greater flexibility to capture at least the part-based variability within a category, which leads to what is effectively a more lax representation of relative part distance and location. This flexibility of part models is exactly what makes them a current favorite in computer vision for recognition applications involving objects with relatively homogenous structure, such as faces. Since many objects may turn out to be face-like (at least in computer image databases) part models perform well in many cases. This is not unlike certain linguistic applications, in which a structurally poor model can successfully account for the majority of corpus sentences, even as they break down dramatically upon encountering rare constructions or long sentences (Fong & Berwick, 2008). The limitations of the part models in comparison to human performance become clear when categories exhibit a high degree of structural variability. In such cases, investigating the boundaries of these categories might provide us with evidence of the representational capacity of the visual system.

Note however, that the more complex model is not always the better model. The supergrammar model would have provided a good fit in the cases of overgeneralization on the part of the subjects. This is why the supergrammar model improves against the grammar model in the absence of a study phase, when the general uncertainty of subjects may lead to incorrect overgeneralization (Figure 4, right). However, in the majority of cases, the supergrammar performs worse than the grammar model. This indicates that subjects were able to zero in not merely on the representation which afforded the greatest complexity, but on the representation which afforded the right degree of complexity for the data at hand.

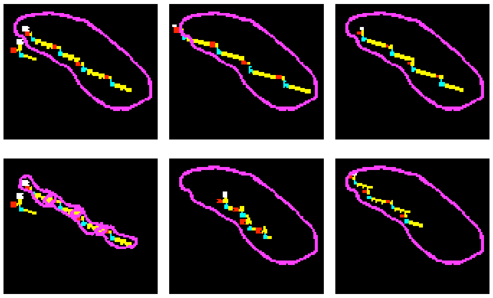

Figure 5: An illustration of the limitations of the part model. Top left: a typical (correct) contour around a recursively constructed object. Bottom left: A typical oversegmentation of the same object. Top middle and right: Explanations of the contour proposed by the grammar model (mean and max likelihood). Bottom middle and right: Explanation of the contour proposed by the part model (mean and max likelihood).

Conclusion

In this paper, we asked to what extent visual representations reflect the complex structural characteristics of some classes of existing visual objects which exhibit recursion and hierarchical organization of unbounded depth. Our results indicate that structurally poor models which allow for some part-based variation capture some but not all of the representational strategies of the visual systems. We find that under certain conditions people exhibit the kind of generalization more consistent with a grammar-based generative model, rather than an exemplar-based matching model. Many questions remain unanswered, most notably, what inference mechanisms are engaged in the process of acquiring such representations from data, how much and what kind of data is necessary, and to what extent everyday visual tasks, such as scene parsing, tap into the full range of available structural representations. Last but not least, the involvement of complex structural generative models in the visual domain raises the tantalizing possibility that representations across different cognitive domains share a single abstract foundation. Further research on the boundaries of human representational capacity promises to shed light on some of these important issues.

References

Bierderman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4).

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. *IEEE Conf. Computer Vision and Pattern Recognition*.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.

Fong, S., & Berwick, R. (2008). Treebank parsing and knowledge of language: A cognitive perspective. *Proceedings of CogSci*.

Fu, K. S. (1974). *Syntactic methods in pattern recognition*. Academic Press.

Hegde, J., Bart, E., & Kersten, D. (2008). Fragment-based learning of visual object categories. *Current Biology*, *18*, 597-601.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure. In *Proceedings of the royal society of london b, vol. 200.*

Palmer, S. (1999). *Vision science: Photons to phenomenology*. Bradford Books, MIT Press.

Savova, V., & Tenenbaum, J. B. (2008). A grammar-based approach to visual category learning. *In Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Wang, W., Pollak, I., Wong, T.-S., Bouman, C., Harper, M., & Siskind, J. (2006). Hierarchal stochastic image grammars for classification and segmentation. *IEEE Transactions on Image Processing (TIP)*, *15*.

Zhu, L., Chen, Y., & Yuille, A. L. (2006). Unsupervised learning of a probabilistic grammar for object detection and parsing. *NIPS*, 1617-1624.