Exemplar Frequency Affects Unsupervised Learning of Shapes

Nathan Witthoft (witthoft@stanford.edu)
Department of Psychology, Jordan Hall, 450 Serra Mall, Building 420
Stanford, CA 94305 USA

Nicolas Davidenko (ndaviden@psych.stanford.edu)
Department of Psychology, Jordan Hall, 450 Serra Mall, Building 420
Stanford, CA 94305 USA

Kalanit Grill-Spector (kalanit@stanford.edu)
Department of Psychology, Jordan Hall, 450 Serra Mall, Building 420
Stanford, CA 94305 USA

Abstract

Exposure to the spatiotemporal statistics of the world is thought to have a profound effect on shaping the response properties of the visual cortex and our visual experience. Here we ask whether subjects' discrimination performance on a set of parameterized shapes changes as a function of the distribution with which the shapes appear in an unsupervised paradigm. During training, subjects performed a fixation task while shapes drawn from a single axis of a parameterized shape space appeared in the background. The frequency with which individual shapes appeared was determined by imposing a normal distribution centered on the middle of the shape axis. Comparison of performance on a shape discrimination task pre and post training showed that subjects' d-prime increased as a function of the frequency with which the exemplars appeared despite the lack of feedback and engagement in a simultaneous task not directed at the shapes. Performance on an untrained set of shapes was largely unchanged across the two testing sessions. This suggests that the visual system may optimize representations by fitting itself to the distribution of experienced exemplars even without feedback, providing the most discriminative power where examples are most likely to occur.

Keywords: Unsupervised learning, vision, perceptual learning.

Background

How people are able to discriminate visually similar items while recognizing the same item across dramatic image transformations is one of the fundamental problems of vision. Experience is thought to play a critical role in forming the underlying cortical representations that support these abilities. One possibility that has been explored in computational and behavioral studies is that the visual system is able to discover and take advantage of statistical regularities in the retinal input via simple unsupervised learning mechanisms (Barlow, 1989a). Our proposal is that unsupervised learning of the frequency of exemplars may fine-tune cortical representations to best match the distribution of exemplars within a category, thus providing the selectivity needed to discriminate between highly similar images where they are most likely to occur.

Unsupervised learning is a process whereby the brain receives inputs but obtains neither supervised target outputs, feedback, nor rewards and as a result finds patterns in the data beyond what would be considered random noise (Ghahramani, 2004). The theoretical framework is based on the notion that the brain's goal is to build representations of the input (even without feedback) that can be used for decision making and predicting future inputs (Poggio et al., 1992). These self-organizing mechanisms could play a crucial role in transforming the continuous flux of retinal stimulation into the stable recognizable objects of our everyday experience. It is important to note that there may be internal reward that guides learning (Seitz and Watanabe, 2005), but this takes place in the absence of explicit feedback on performance.

Numerous studies have shown that the visual system adjusts itself as a function of experience even in situations where subjects are uninstructed. Adaptation represents a phenomenon of this kind, where prolonged exposure to some stimulus value can shift the sensitivity of the visual system for a short period of time. For example, viewing rightward motion causes subsequently presented static stimuli to appear as though they are moving to the left (Anstis et al., 1998). Such aftereffects are perceptually compelling, and can be found for a wide variety of visual features ranging from the relatively simple such as line orientation to the very complex such as facial identity (Leopold et al., 2001; Witthoft et al., 2006) and do not require instruction or feedback (though some may require attention; Moradi et al., 2005). With respect to our proposal, it has been argued that adaptation is not just a useful way for psychologists to probe the visual system, but reflects a functional mechanism by which vision increases its sensitivity to changes in recent experience (Webster et al., 2001; Barlow & Foldiak, 1989; Clifford & Rhodes, 2005).

Studies of perceptual learning also show experience dependent changes, but have often relied on the notion that improvement in discrimination is heavily task dependent and that mere exposure may not be enough to drive performance changes (Shiu & Paschler, 1992; Karni & Sagi, 1991). However, recent work has suggested discrimination performance can improve as a result of unsupervised learning and when subjects are engaged in an orthogonal task. For example, Watanabe et al. (Nature, 2001) had subjects perform a central letter detection task superimposed on a field of moving dots. 5% of the dots were moving coherently in the same direction, but this coherence was below the subjects' detection threshold meaning they were both unable to detect the direction of coherence and engaged in a different task. Following training, subjects showed improved performance in direction discrimination of just above threshold coherence moving dots, but only in the vicinity of the direction seen during training. The authors argued that this irrelevant learning (as they call it) means the visual system can increase its sensitivity to a frequently occurring feature or stimulus even when it is not task relevant. More recently, the same authors have suggested that internal feedback is being supplied when there is a success on the orthogonal task, and that at that time both task relevant and task irrelevant features are strengthened (Seitz and Watanabe, 2005).

Lastly, numerous experiments have shown that subjects can successfully learn the spatio-temporal structure of visual stimuli even in the absence of explicit instruction or feedback. For example, following training on sequences of shapes in which triplets repeat, subjects are able to discriminate seen before triplets from novel ones (Fiser & Aslin, 2002; Turk-Brown et al., 2005). Similar results have been obtained for arrangements of shapes within an image (Fiser & Aslin, 2001) and for sequences using natural scenes (Brady & Oliva, 2008) that show transfer to sequences that only share categorical similarity and words. Other work has shown that invariant object recognition (the ability to recognize exemplars as the same across image transformations such as position or view) may result from the visual system taking advantage of spatio-temporal correlations in the input (Cox et al., 2005; Wallis & Bulthoff, 2001; Sinha & Poggio, 1996) While our goal is not to evaluate subjects' ability to do the kind of visual statistical learning for the generation of invariances, these studies do show that subjects are sensitive to spatio-temporal information in the visual input.

The idea that visual discrimination and recognition might be shaped by the distribution of experienced items has been examined most extensively in the face domain. Some models of face perception (Valentine, 1991) propose that faces are represented as points in an abstract vector space, centered on the norm face representing the average of the faces a person has seen (Leopold et al., 2001; Rhodes & Jeffrey, 2006). It is typically assumed that faces are normally distributed around the mean face with the highest concentration of faces occurring near the norm. Clearly the diet of faces that a particular individual is exposed to is heavily dependent on their local environment and it is believed that the set of features that might form an individual's space are tuned by that person's experience.

An experiment that relies on training with faces to test this idea may suffer from competition with the overwhelming experience a subject brings to the laboratory. However, a clever experiment by Webster et al. (Nature 2004) navigated this obstacle by asking Asian and Caucasian-American subjects to set an ethnicity boundary on a set of morphed faces between an Asian and a Caucasian face. Each group set that boundary closer to their own ethnicity, presumably closer to the average of faces they had seen and reflecting an increased sensitivity to deviations from their own experience. Interestingly, a second group of Asian subjects who had been in the United States for approximately one year also did the task, and their boundary was shifted towards the Caucasian end of the dimension with the size of the shift significantly correlated with how long they had been in the US. While this is not a direct test of the hypothesis, as the experimenters could not manipulate the frequencies of exposure across face space directly, it is suggestive that subjects' representation of faces is driven by the distribution of the input and that such changes can occur even in adults.

While there has been much debate over whether the mechanisms serving face perception are distinct from those for other kinds of objects (Farah et al., 1998), it seems likely that mechanisms that adjust sensitivity in order to match the distribution of similar exemplars would be generally useful across all classes of objects. To test this idea we created sets of novel parameterized shapes, allowing us to control stimulus variation and frequency that subjects would have no experience with, thus allowing a direct test of the role of frequency in determining sensitivity. One way to view our shape dimensions is that each dimension represents a category where the within category exemplars vary along a physical continuum (as with faces). In our preliminary experiment, subjects' ability to discriminate similar shapes at various points along a shape dimension was tested before and after unsupervised training. During training we manipulated the frequency with which subjects experienced different parts of the shape dimension by sampling shapes from a normal distribution centered on the middle of the shape axis. Our hypothesis is that subjects will become most sensitive to changes in the regions of the shape dimension that are most frequently experienced even in the absence of feedback and while they are engaged in an orthogonal visual task. This change in sensitivity would represent one way that the visual system could adapt to environmental statistics to produce useful representations for perception (Goldstone, 1998, Barlow & Foldiak, 1989, Clifford & Rhodes, 2005).

Stimuli

Four "prototype shapes" were created using MATLAB. Each prototype shape was constructed by placing 16 keypoints on a grid, plus 3 additional points that were fixed across prototypes (Figure 1A). The points were made to correspond across all prototypes, such that the 5th point on one prototype corresponded to the 5th point on all other prototypes, etc. The points were connected (Figure 1B) and these connections were then replaced by smooth bi-cubic splines (Figure 1C). Finally, the resulting closed shape was filled in (Figure 1D). Care was taken when placing keypoints to avoid loops in the final shape.

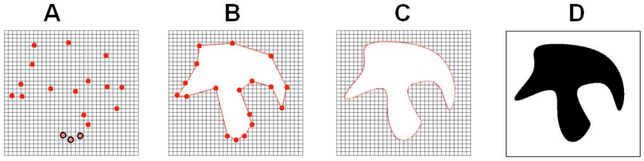

Figure 1. The construction of a prototype shape. (A) Keypoints (16 variable and 3 fixed (outlined)) are placed on a grid; (B) the points are connected; (C) the connections are replaced with smooth bi-cubic splines; (D) the resulting shape is filled in.

From the 4 prototype shapes (see Figure 2), two continuous axes were generated joining two pairs of prototypes. Since each shape can be fully represented by the positions of its 16 variable keypoints, intermediate shapes between two prototypes can be formed by linearly weighting the positions of those keypoints across the prototypes and recomputing the splines. For example, for an intermediate shape 25% of the way between Prototype 1 and Prototype 2, the x-position of first keypoint ($x1_{\mathrm{intermediate}}$) would be equal to $.75 * x1_{\mathrm{Prototype1}} + .25 * x1_{\mathrm{Prototype2}}$, and so forth. Figure 2 shows the two axes used in our training study, each consisting of 7 reference shapes. For the discrimination tasks, 4 additional shapes were created for each reference shape, 2 on each side of the shape axis. One pair was nearer and considered the hard discrimination, and one pair was farther and considered the easy discrimination.

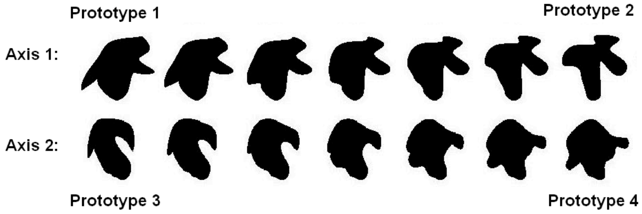

Figure 2. Two axes of shapes used in our training study. The intermediate shapes between the prototypes (endpoints) are constructed by linearly weighting the positions of the keypoints and then recomputing the splines.

Methods

Subjects

8 subjects (4 female) participated in the experiment in exchange for payment. All the experiments were programmed using MATLAB and the psychophysics toolbox (Brainard, 1997). The experimental paradigm was approved by the Stanford Human Subjects IRB. All subjects gave informed consent to participate and were paid $10 for each session ($50 total for completing all sessions). Stimuli were presented on Macintosh iMac computers. One subject (female) was removed due to poor performance on the fixation task during training. Including this data point actually increases the significance of all our reported findings, but we cannot be sure that this subject was not attending to the shapes.

Pre-training Discrimination Test

To assess baseline performance, subjects performed a same/different task on pairs of shapes (reference shape and comparison shape). The reference shapes were the 7 examples for each axis shown in Figure 2. The comparison shapes were taken from either side of each reference, with one closer pair (hard discrimination) and one farther pair (easier discrimination). We used two levels of difficulty as we were uncertain as to how initial discriminability performance might interact with training (Watanabe et al., 2001). On each trial subjects were presented with a shape for 200 milliseconds, followed by a 500 millisecond delay, and then a second shape for 200 milliseconds. Subjects were instructed to indicate whether the two shapes were the same or different by pressing the appropriate key. A new trial was not initiated until the subject responded. For each reference shape there were two levels of comparison trials (easy and hard) and two sides of the shape axis from which comparison shapes were drawn. Each possible comparison was measured 15 times for a total of (4 comparisons x 7 shapes x 15 times x 2 shape dimensions) 840 different trials. For each of the 14 shapes there were 30 same trials. Trials were presented in a random order and no feedback was given on any of the trials. Subjects were tested on shapes from both shape dimensions, but only shapes from axis 1 were shown during training.

Unsupervised Training

Unsupervised training took place on each of the 3 days following the initial discrimination experiment. Subjects performed a fixation change-detection task, and were instructed to press a key when the fixation cross slightly changed in size. The fixation cross was superimposed on a stream of shapes that appeared at 2 Hz (the shape presented for 300 milliseconds followed by a 200 millisecond blank) for the duration of the experiment. These shapes were sampled with a normal distribution discretized into 73 intervals centered on the shape dimension used in training. 6000 shapes were shown in each session that lasted approximately one hour each day for a total of 18000 shapes and 3 hours of training. Each day's training was divided into 6 blocks of 1000 trials. There were 40 fixation changes for each block. Timing of the fixation changes was set by dividing each block into 40 equal intervals and then inserting one change at a random time-point in each interval. Between blocks subjects were permitted to pause as long as they felt they needed to. Presentation of the stimuli was such that the distribution of shapes was normal at the block level. Critically, subjects were not instructed to attend to the shapes nor were they given feedback of any kind. The fixation task was intended to draw attention away from the shapes thus testing for task irrelevant learning (Seitz and Watanabe, 2005).

Post-training Discrimination Test

The post training discrimination test used the same stimuli and design as the pre-training test and was conducted on the day after the last training day.

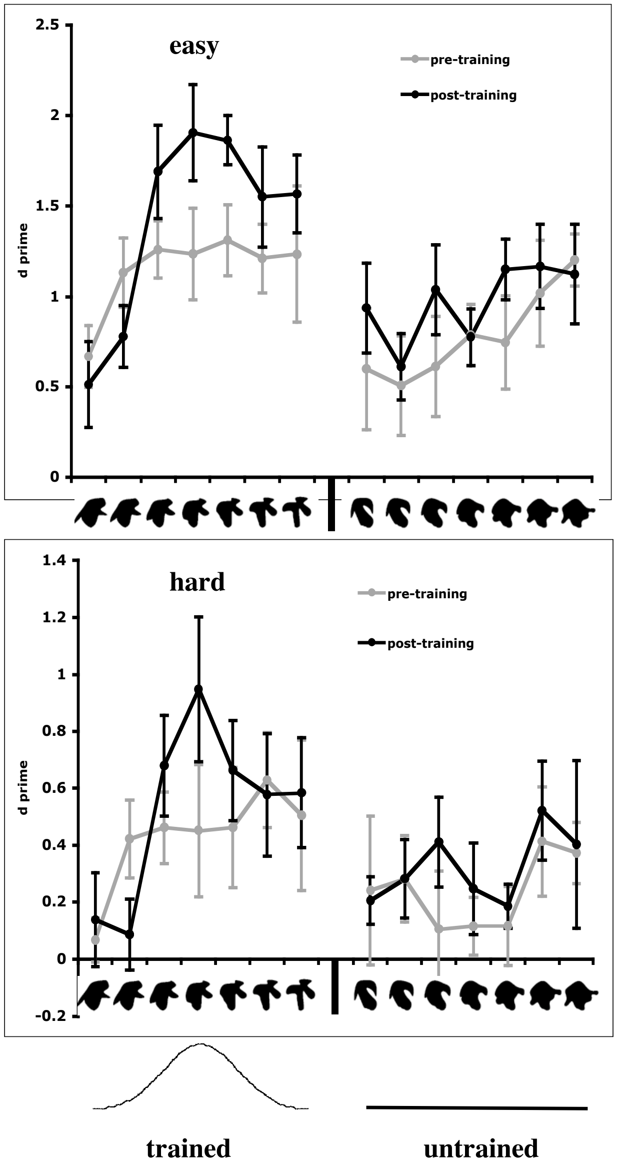

Figure 3. Performance of subjects on shape discriminations before and after training. The top panel shows the easy discrimination and the bottom the hard discrimination. Note the y axes differ slightly, in order to show the effect for each level of discrimination. The x-axis shows the reference shapes used in the experiment. At bottom are qualitative graphics indicating the distribution of shapes during training. For the trained it is Gaussian centered on the shape axis, and for the untrained it is flat and at 0 (i.e. no training).

Results

Examination of performance on the fixation task during the training phase of the experiment showed that subjects attended to fixation, detecting changes in cross size 91 percent of the time (SE= 2 percent).

For each subject false alarm and hit rates were separately calculated for each comparison and reference stimulus for the hard and easy conditions. The hits and false alarm rates were used to calculate d-primes(Green & Swets, 1966). Separate 2 (pre vs post training) by 7 (reference shape) repeated measures ANOVAs were used to examine the effects of training in the 4 conditions (hard or easy discrimination on the trained or untrained shape dimension). Training showed similar but not identical effects for both the hard and easy discriminations (Figure 3 left, bottom and top panels respectively). For the hard discriminations, there was a main effect of training, with higher d-prime following training (F(1,13)=10.5, $p<0.05$). There was also an effect of stimulus (F(5,13)=2.9, $p<0.05$) reflecting the fact that on average, shapes on one end of the dimension were more discriminable than at the other despite the matched distances in the parameterization. For the easy discriminations, the effects of training and position on the stimulus axis were only marginal, (F(1,13)=4, p=0.09 for the training and F(5,13)=8.5, $p<0.05$ for the effect of stimulus).

For the untrained stimuli, the harder discriminations showed no significant improvement in performance post- vs. pre-training (F(1,13)=2.9, $p>0.1$) and no stimulus effect (F(6,13)=1.95, $p>0.1$) (Figure 3 bottom right). In the easy discrimination, subjects did significantly improve as a function of doing the task twice (F(1,13)=6.5, $p>0.05$, Figure 3 top right). There was also a stimulus effect, showing that for the easy discriminations at least, the shapes at one end were easier than shapes at the other. It is important to note that for a direct comparison of the trained vs. untrained conditions a second set of subjects is required who have the relationship between shape dimensions and training swapped. Such an experiment is planned, but here we note that there does not appear to be any effect of learning which is dependent on the position of the stimulus on the dimension.

A correlation analysis was used to more directly assess the relationship between changes as a result of training and the frequency with which each stimulus appeared during training. For each prototype, we created a normalized frequency score by dividing the number of times a reference shape appeared by the maximum number of times a shape could appear (i.e. how often the shape at the center of the distribution appeared). Critically, the mean improvement in d-prime was predicted by the frequency of stimuli during training for the hard discrimination, with stimuli that were seen more often during training producing a larger mean improvement in d-prime (r=0.78, $p<0.05$, Figure 4 left). A similar but marginally significant relationship was found for the easy discriminations (r=0.72, p=0.068, Figure 4 right).

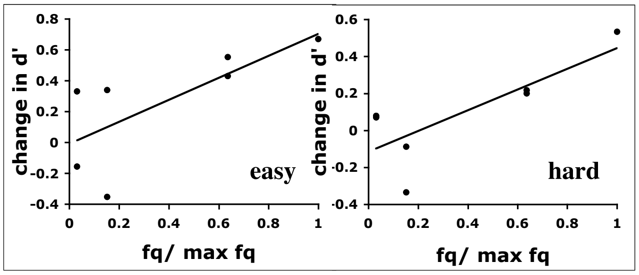

Figure 4. Correlations between normalized frequency and the difference in d-prime pre and post training. Left panel shows the easy discriminations, right panel shows the hard discriminations.

Discussion

These preliminary data show that discrimination performance selectively improved as a function of the frequency with which parameterized shapes were presented in an unsupervised setting while subjects were engaged in another task. Our shape dimensions are meant to be analogous to a within-category set of highly similar exemplars which are encountered with varying frequency. Our results suggest that the visual system automatically tunes itself to this distribution, placing resources where they are most likely to be needed.

Some caveats obviously obtain. First it may be that the learning effects are driven only by frequency rather than the distribution (or relative frequency). One can imagine an experiment with much more training using the same distribution such that the stimuli at the ends of the shape dimension are seen as many times as the shapes in the center of the distribution were seen in the experiment presented here. In the strong version of a matching hypothesis, the same pattern of results would be seen, with little learning at the ends of the shape dimension and improved performance appearing only in the vicinity of the most frequently seen shapes. If learning depends only on the frequency, then improvement should be seen across the entire dimension, possibly reaching an asymptote. Another possibility is that discrimination centers around the mean of the distribution which is here confounded with the frequency. This alternative can be tested by using a heavily skewed distribution and seeing whether the best performance follows the mean or the mode.

Another possibility is that learning is driven by the co-occurrence of detected fixation changes with the shapes. As suggested by Seitz and Watanabe (2005), it may be that detection of a fixation change generates an internal reward, and given that the task is not too demanding, all features present at the time (including the shape onscreen) are reinforced. This intriguing hypothesis can be tested in a number of ways using this paradigm, for example, by only having fixation changes when an infrequently seen shape is present.

Finally, this paradigm offers an interesting way to examine questions of categorization. For example, suppose during training that the shapes are sampled using a bimodal distribution (analogous to say male and female faces). Although shapes that lie between the two modes are presented with much lower frequency, this part of the shape dimension has importance as it could correspond to a natural category boundary (Rosenthal et al., 2001). If the visual system is sensitive to this complex distributional information, an additional increased sensitivity may be found at the boundary even though examples are infrequently presented. Such a result would suggest a second mechanism which is takes advantage of multimodal distributional information to create between category separation.

References

Anstis, S., Verstraten, F. A. J., & Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences, 2(3), 111-117.*

Barlow, H. B., & Földiák, P. (1989). Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, & G. Mitchison (Eds.), *The Computing Neuron*. Wokingham, England: Addison-Wesley.

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

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes. *Psychological Science, 19(7), 678-685.*

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10, 433-436.*

Clifford, C. W. G., & Rhodes, G. (Eds.) (2005). *Fitting the mind to the world : adaptation and after-effects in high-level vision* (1st ed.). Oxford ; New York: Oxford University Press.

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). 'Breaking' position-invariant object recognition. *Nature Neuroscience,8, 1145-1147.*

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. M. (1998). What is "special" about face perception? *Psychological Review, 105(3), 482-498.*

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

Fiser, J., & Aslin, R. (2002) Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28(3), 458-467.*

Ghahramani, Z. Unsupervised Learning. In Bousquet, O., von Luxburg, U. & G. Raetsch (Eds.) *Advanced Lectures* *in Machine Learning. Lecture Notes in Computer Science.* Springer-Verlag, Berlin, 2004.

Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology, 49, 585-612.*

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory.* New York: Wiley.

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

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4(1), 89-94.*

Moradi, F., Koch, C., & Shimojo, S. (2005). Face adaptation depends on seeing the face. *Neuron, 45, 169-175.*

Poggio, T., Fahle, M. & Edelman, S. (1992) Fast perceptual learning in visual hyperacuity. *Science 256, 1018-21.*

Rhodes, G., & Jeffrey, L. (2006). Adaptive norm based coding of facial identity. *Vision Research, 46, 2977-2987.*

Rosenthal, O., Fusi, S., & Hochstein, S. (2001) Forming classes by stimulus frequency: behavior and theory. *Proceedings of the National Academy of Sciences, 98, 4265-4270.*

Seitz, A. & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences, 9(7), 329-334.*

Shiu, L. & Pachler, H. (1992) Improvement in line orientation discrimination is retinally local but dependent on cognitive set. *Perception and Psychophysics, 52(5), 582-588.*

Sinha, P., & Poggio, T. (1996). Role of learning in three dimensional form perception. *Nature, 384, 460-463.*

Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134, 552-564.*

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43, 161-204.*

Wallis, G., & Bulthoff, H. H. (2001). Effects of temporal association on recognition memory *Proceedings of the National Academy of Sciences, 98(8) 4800-4804.*

Watanbe, T., Nanez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature, 413, 844-848.*

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, 428(6982), 557-561.

Witthoft, N., Winawer, J., & Boroditsky, L. (2006) How looking at someone you don't know can help you recognize someone you do. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*