Probabilistic principles in unsupervised learning of visual structure: human data and a model

Shimon Edelman, Hwajin Yang
& Benjamin P. Hiles
Department of Psychology
Cornell University, Ithaca, NY 14853
{*se37,hy56,bph7*}*@cornell.edu*

Nathan Intrator
Institute for Brain and Neural Systems
Box 1843, Brown University
Providence, RI 02912
Nathan_Intrator@brown.edu

Abstract

To find out how the representations of structured visual objects depend on the co-occurrence statistics of their parts, we exposed subjects to a set of composite images with controlled conditional probabilities of the constituent fragments. We then compared the part verification response times for various probe/target combinations before and after the exposure. With composite probes, the drop in verification RT following exposure was much larger for targets that contained pairs of fragments perfectly predictive of each other, compared to those that did not; for lone-fragment probes, this difference was reversed. This pattern of results is consistent with the principle according to which objects should be treated as wholes, unless their parts are observed sufficiently frequently in more than one configuration.

1 Motivation

How does the human visual system decide for which objects should it maintain distinct and persistent internal representations of the kind typically postulated by theories of object recognition? Consider, for example, the image shown in Figure 1, left. This image can be represented as a monolithic hieroglyph, a pair of Chinese characters (which we shall refer to as A and B), a set of strokes, or, trivially, as a collection of pixels. Note that the second option is only available to a system previously exposed to various combinations of Chinese characters. Indeed, the decision whether to represent this image as $\{AB\}$, $\{A, B\}$, $\{A, B, AB\}$ or otherwise can only be made in a principled manner on the basis of prior exposure to related images.

According to Barlow's [1] insight, the tally of *suspicious coincidences* offers a useful principle: two candidate fragments A and B should be combined into a composite object AB if the probability of their joint appearance $P(A, B)$ is much higher than $P(A)P(B)$, which is the probability expected in the case of their statistical independence. This criterion may be compared to the Minimum Description Length (MDL) principle, which has been previously discussed in the context of object representation [2, 3]. In a simplified form [4], MDL calls for representing AB explicitly as a whole if $P(A, B) \gg P(A)P(B)$, just as the principle of suspicious coincidences does.

While the MDL criterion $P(A, B)/\left(P(A)P(B)\right)$ certainly indicates a suspicious coincidence, we believe that additional probabilistic considerations may be used in setting the degree of association between A and B. One example is the possible perfect predictability of A from B and vice versa, as measured by $minCP \doteq \min\{P(A|B), P(B|A)\}$. If $minCP = 1$, then A and B are perfectly predictive of each other and should really be coded by a single symbol, whereas the MDL criterion may suggest merely that some association between the representation of A and that of B be established. In comparison, if A and B are *not* perfectly predictive of each other ($minCP < 1$), there is a case to be made in favor of coding them separately to allow for a maximally expressive representation, whereas MDL may actually suggest a high degree of association (if $P(A,B)/\left(P(A)P(B)\right) \gg 1$). In this study we investigated whether the human visual system uses a criterion based on $minCP$ alongside MDL while learning to represent composite objects.

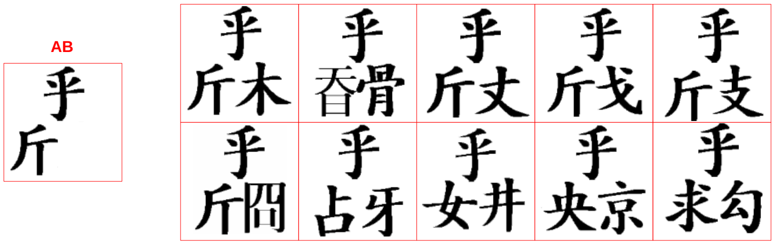

Figure 1: *Left:* How many objects are contained in image AB? Without prior knowledge, a reasonable answer, which embodies a holistic bias, should be “one.” *Right:* In this set of ten images, AB appears five times as a whole; the other five times a fragment wholly contained in AB appears in isolation. This statistical fact provides grounds for considering AB to be composite, consisting of two fragments (call the upper one A and the lower one B), because $P(A|B) = 1$, but $P(B|A) = 0.5 < 1$.

To date, psychophysical explorations of the sensitivity of human subjects to stimulus statistics tended to concentrate on means (and sometimes variances) of the frequency of various stimuli. One recent and notable exception is the work of Saffran et al. [5], who showed that infants (and adults) can distinguish between “words” (stable pairs of syllables that recur in a continuous auditory stimulus stream) and non-words (syllables accidentally paired with each other, the first of which comes from one “word” and the second – from the following one). Thus, subjects can sense (and act upon) differences in transition probabilities between successive auditory stimuli. In the visual domain, Fiser and Aslin [6] presented subjects with geometrical shapes in various configurations, and found effects of conditional probabilities of shape co-occurrences, in a task that required the subjects to decide in each trial which of two simultaneously presented shapes was more familiar.

The present study was undertaken to investigate the relevance of the various notions of statistical independence to the unsupervised learning of complex visual stimuli by human subjects. Our experimental approach differs from that of [6] in several respects. First, instead of explicitly judging shape familiarity, our subjects had to verify the presence of a probe shape embedded in a target. This objective task, which produces a pattern of response times, is arguably better suited to the investigation of internal representations involved in object recognition than subjective judgment [7]. Second, the estimation of familiarity requires the subject to access in each trial the representations of all the objects seen in the experiment; in our task, each trial involved just two objects (the probe and the target), potentially sharpening the focus of the experimental approach. Third, with the subjects’ sensitivity to conditional probabilities demonstrated by Fiser and Aslin, we decided to concentrate on specific predictions generated by the various notions of stimulus independence, such as MDL and $minCP$.

2 The experiment

Subjects were presented with two blocks of yes/no trials (“is the probe contained in the target?”; cf. [7]), one before and the other after exposure to a set of stimuli composed of Chinese characters such as those in Figure 1, right. The conditional probabilities of the appearance of individual characters were controlled. Two characters A, B could be *paired*, in which case we had $P(A|B) = P(B|A) = 1$. Alternatively, A, B could be *unpaired*, with $P(A|B) = 1$, $P(B|A) = 0.5$. In either case, we had $P(A,B)/\left(P(A)P(B)\right) \approx 8$. Thus, for *paired* A, B the minimum conditional probability $minCP = \min\left\{P(A|B), P(B|A)\right\} = 1$ and the two characters were perfectly predictable from each other, whereas for *unpaired* A, B $minCP = 0.5$, and they were not. In the latter case AB probably should not be represented as a whole. Would the subjects’ behavior reflect the use of this, rather extreme, criterion of independence, or would they employ a criterion related more closely to the principles of suspicious coincidences or MDL?

Let us assume that the subjects tally the conditional probabilities of various pairings of potential standalone fragments [6], and, furthermore, that they maintain explicit and persistent representations for fragments or for fragment groups, as suggested by the independence criteria. Such representations should then support a kind of priming [8]: the response time in trials in which the probe is explicitly represented should be faster than in trials in which the probe is represented in a distributed fashion. The representations, however, can only be acquired by the subjects through the process of assimilating the training set. This leads to the main prediction of the study: the subjects should respond faster after exposure to the training set than before — but only in those trials in which the probe (embedded in the target; see Figure 2, left) is assigned a separate and explicit representation. The experiment, therefore, hinges on a comparison of the patterns of verification response times before and after exposure to the training set.

The experiment involved two types of probe conditions: PTYPE=`Fragment`, or $A \rightarrow ABZ$ (with $V \rightarrow ABZ$ as the reference condition), and PTYPE=`Composite`, or $AB \rightarrow ABZ$ (with $VW \rightarrow ABZ$ as reference). In this notation (see Figure 2, left), A and B are “familiar” fragments with controlled minimum conditional probability $minCP$, and V, W, X are novel random fragments. The experiment consisted of a baseline session, followed by training exposure (unsupervised learning), followed in turn by the test session (Figure 2, right). In the baseline and test sessions, the subjects had to indicate whether the probe was contained in the target. In the training session, the subjects had to note the order in which the three characters appeared on the screen. Fourteen subjects, none of them familiar with the Chinese writing system, participated in the experiment in exchange for course credit.

3 Results

We carried out a mixed-effects repeated measures analysis of variance (SAS procedure MIXED [9]) for ΔRT and SPEED-UP, with PTYPE and $minCP$ as independent variables. The dependent variables are defined and the results are summarized in Figure 3.

The effect of training on SPEED-UP (Figure 3, top left panel) was strikingly different for composite probes with $P = 1$ (404 ms) compared to the other three conditions (170 ms on the average). ANOVA revealed significant main effects of PTYPE ($F_{1,13} = 5.31, p < 0.04$), $P(A,B)$ ($F_{1,13} = 6.05, p < 0.03$) and the interaction ($F_{1,13} = 6.59, p < 0.02$).

An analysis of the ΔRT data revealed that subjects’ response was, on the average, slowed down by composite probes (mean $\Delta RT = -93$ ms), compared to fragment probes (89 ms). In other words, it took relatively much longer to determine correctly that the probe was contained in the target (compared to the time to determine that it was not) if the

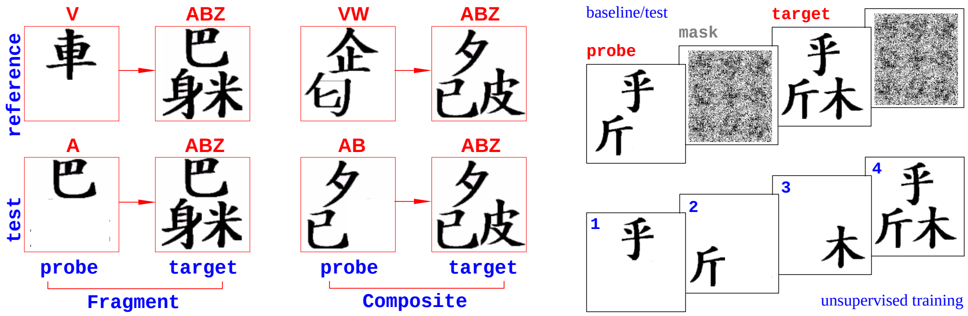

Figure 2: *Left:* Illustration of the probe and target composition for the two levels of PTYPE (`Fragment` and `Composite`). For convenience, the various categories of characters that appeared in the experiment are annotated here by Latin letters. Specifically, A, B stand for characters with controlled $minCP = \min\{P(A|B), P(B|A)\}$; the training set was constructed with $minCP = 0.5$ for some pairs, and $minCP = 1$ for others. V, W, Z stand for characters that appeared only once throughout the experiment. In negative or `reference` trials, the correct answer was *no* (i.e., probe not contained in target); in positive or `test` trials, the correct answer was *yes*. ΔRT was defined as $RT(\texttt{reference}) - RT(\texttt{test})$, which makes it positive if the response time is shorter in the test condition relative to the reference condition. *Right Top:* The structure of a part verification trial (same for baseline and test phases). The probe stimulus was followed by the target (each presented for 150 ms; a mask was shown before and after the target). The subject had to indicate whether or not the former was contained in the latter (in this example, the correct answer is *yes*). A sequence consisting of 64 trials like this one was presented twice: before training (baseline phase) and after training (test phase). *Right Bottom:* The structure of a training trial (the training phase, placed between baseline and test, consisted of 80 such trials). The three components of the stimulus appeared one by one for 150 ms to make sure that the subject attended to each, then together for 700 ms. The subject was required to note whether the sequence unfolded in a clockwise or counterclockwise order.

probe was a composite object. It is not this bias against composite probes, however, but the *differential* effect of training on ΔRT in the four conditions (two levels of PTYPE × two levels of $minCP$) that is relevant to testing the $minCP$ hypothesis. Indeed, training precipitated a drastic change in the effect of composite probes with $minCP = 1$, from -250 ms in the baseline phase to 7 ms in the test phase (Figure 3, top middle and right); no such change was found in the other conditions.[1] Note that it is in this condition that the two constituents of the target are perfectly predictable from each other, providing the subjects' visual system with the greatest incentive to form a unified representation of the target, and thereby boosting the effect of composite probes.

This behavior conforms to the predictions of the $minCP$ principle: subjects seem to have represented *paired* characters together, while splitting apart *unpaired* ones. Note that the suspicious coincidence ratio was the same in both cases, $r_{susp} \doteq P(A,B)/\left(P(A)P(B)\right) \approx 8$. Thus, the subjects in this experiment proved to be sensitive to the $minCP$ measure of independence, over and above the (constant-valued) MDL-related criterion, according to which the propensity to form a unified representation of two fragments, A and B, should depend on r_{susp} [1, 4].

[1] This observation is supported by the ANOVA, which revealed a highly significant PHASE × PTYPE × $minCP$ interaction ($F_{1,13} = 11.71, p < 0.0045$). The other significant effects were PTYPE ($F_{1,13} = 38.88, p < 0.0001$), PHASE × PTYPE ($F_{1,13} = 9.27, p < 0.0094$), and a marginal PHASE × $minCP$ interaction ($F_{1,13} = 3.53, p < 0.08$).

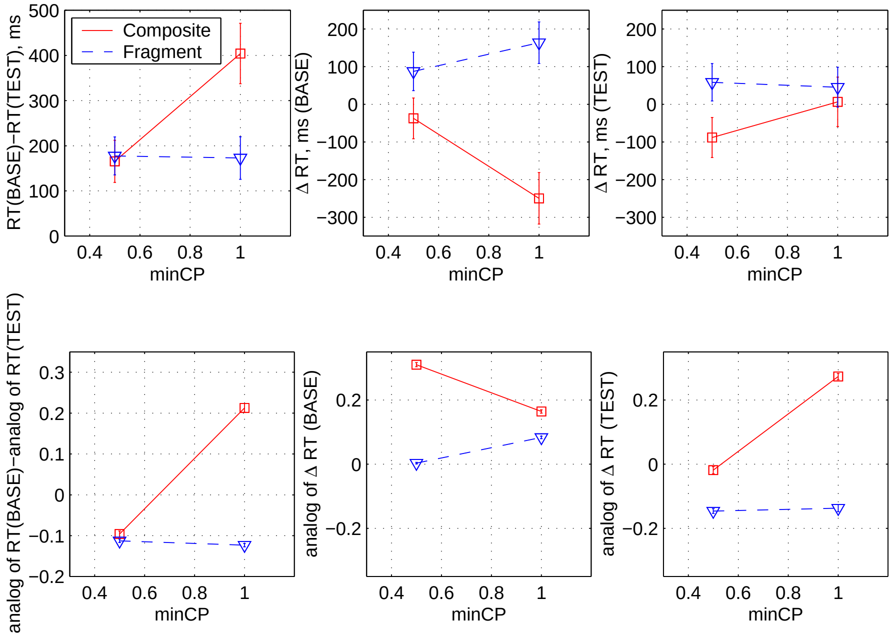

Figure 3: **Top row:** Results from human subjects. The effects of training on RT and on ΔRT (least-squares estimates of means and standard errors, computed by the LSMEANS option of SAS procedure MIXED), plotted vs. $minCP$, by probe type. *Left:* SPEED-UP, defined as the difference in RT between baseline and test phases. The SPEED-UP for composite probes (solid line) with $minCP = 1$ exceeded that in the other conditions by more than 200 ms. *Middle and Right:* ΔRT, defined as the difference in RT between negative and positive trials for each condition; the two plots correspond to baseline and test phase, respectively. Following training, ΔRT was significantly boosted for composite probes and somewhat reduced for fragment probes for $minCP = 1$; no such effect was found for $minCP = 0.5$. All these findings are consistent with the notion that exposure to character pairs with $minCP = 1$, but not to those with $minCP = 0.5$, caused subjects to treat the members of a pair as a single object by forming an integral representation of the two; the existence of such representations manifested itself in an increased speed-up effect for composite, but not fragmented, probes. **Bottom row:** Results of a simulated experiment, in which a model outlined in section 4 was given the same 80 training images as the human subjects. The difference of reconstruction errors for probe and target served as the analog of RT; baseline measurements were conducted on half-trained networks. The error bars are smaller than the symbols in this plot.

We are currently studying the effects of varying r_{susp} independently of $minCP$. Because of the nature of these variables, a mixed within- and between-subjects design must be used, which requires a large number of subjects. Preliminary results involving all four combinations of $r_{susp} \in \{1.13, 8.33\}$ and $minCP \in \{0.5, 1.0\}$, obtained with 17 subjects (about four subjects per condition) indicate that the effects of $minCP$ found in the first experiment are much stronger for $r_{susp} = 8.33$ than for $r_{susp} = 1.13$. This suggests that the influence of the two criteria, $minCP$ and r_{susp}, on the representation of composite objects is synergistic.

4 An unsupervised learning model and a simulated experiment

The ability of our subjects to construct representations that reflect the probability of co-occurrence of complex shapes has been replicated by a novel unsupervised learning model, described elsewhere [10]. The model (Figure 4) is based on the following observation: an auto-association network fed with a sequence of composite images in which some fragment/location combinations are more likely than others develops a non-uniform spatial distribution of reconstruction errors; smaller errors appear in those locations where the image fragments recur. This information can be used to form a spatial receptive field for the learning module, while the reconstruction error can signal its relevance to the current input [11, 12]; different modules learn to represent different "what+where" combinations through competition. The performance of this model, trained on precisely the same sequence of triplets of Chinese characters as our human subjects, is shown in Figure 3, bottom row. The differential effects of $minCP$ for the two probe kinds (`Fragment` and `Composite`) are the same for humans and for the model.[2]

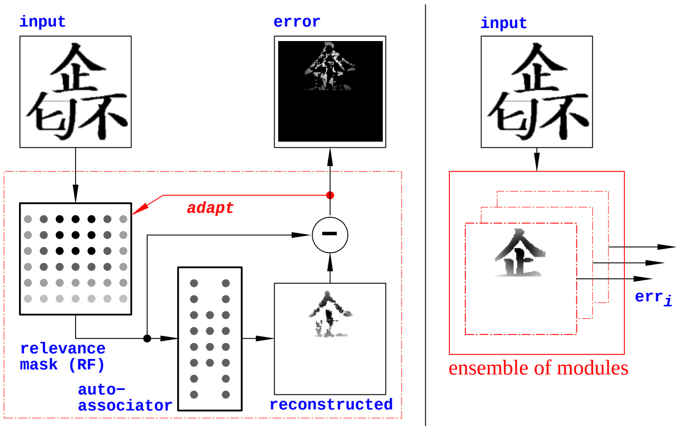

Figure 4: *Left:* Functional architecture of a fragment module. The module consists of two adaptive components: a reconstruction network, and a relevance mask, which assigns different weights to different input pixels. The mask modulates the input multiplicatively, determining the module's receptive field. Given a sequence of images, several such modules working in parallel learn to represent different categories of spatially localized patterns (fragments) that recur in those images. The reconstruction error serves as an estimate of the module's ability to deal with the input ([11, 12]; in the error image, shown on the right, white corresponds to high values). *Right:* The Chorus of Fragments (CoF): a bank of fragment modules, each tuned to a particular shape category, appearing in a particular location [4]. The unsupervised competitive procedure used to learn the representations of the various fragments ("what") and the corresponding mask weights ("where") is described in [10].

5 Discussion

Human subjects have been previously shown to be able to acquire, through unsupervised learning, sensitivity to transition probabilities between syllables of nonsense words [5] and between digits [13], and to co-occurrence statistics of simple geometrical figures [6]. Our

[2] The pattern of mean "RT" produced by the model lacks the bias in favor of fragment probes exhibited by humans; this effect is orthogonal to the issue of probability tallying, and is, therefore, outside the scope of the present work.

results demonstrate that subjects can also learn (without awareness; cf. [13]) to treat combinations of complex visual patterns differentially, depending on the conditional probabilities of the various combinations, accumulated during a short unsupervised training session.

In the present study, the criterion of suspicious coincidence between the occurrences of A and B is met in both $P(A|B) = 0.5$ and $P(A|B) = 1$ conditions: in each case, $P(A,B)/\left(P(A)P(B)\right) \approx 8$. Yet, the subjects' behavior indicates a significant holistic bias: the representation tends to be monolithic (AB), unless imprefect mutual predictability of the potential fragments (A and B) provides support for representing them separately. A similar holistic bias, operating in a setting where a single encounter with a stimulus can make a difference, is found in language acquisition: an infant faced with an unfamiliar word will assume it refers to the *entire shape* of the most salient object [14].

The computationally challenging unsupervised learning task faced by our subjects (and our model) can be addressed using information-theoretic and probabilistic methods [15, 16], including MDL [2]. Our current research focuses on further elucidation of the manner in which subjects process statistically structured data, and on development of the new model of structure learning outlined in the preceding section [10].

References

[1] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[2] R. S. Zemel and G. E. Hinton. Developing population codes by minimizing description length. *Neural Computation*, 7:549–564, 1995.

[3] E. Bienenstock, S. Geman, and D. Potter. Compositionality, MDL priors, and object recognition. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Neural Information Processing Systems*, volume 9. MIT Press, 1997.

[4] S. Edelman and N. Intrator. A productive, systematic framework for the representation of visual structure. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 10–16. MIT Press, 2001.

[5] J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928, 1996.

[6] J. Fiser and R. N. Aslin. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, -:–, 2001. in press.

[7] S. E. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9:441–474, 1977.

[8] C. L. Wiggs and A. Martin. Properties and mechanisms of perceptual priming. *Curr. Opin. Neurobiol.*, 8:227–233, 1998.

[9] SAS. *User's Guide, Version 8*. SAS Institute Inc., Cary, NC, 1999.

[10] N. Intrator, S. Edelman, B. P. Hiles, and H. Yang. Unsupervised learning of 'what+where' representations from probabilistic cues, 2001. in preparation.

[11] D. Pomerleau. Input reconstruction reliability estimation. In C. L. Giles, S. J. Hanson, and J. D. Cowan, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 279–286. Morgan Kaufmann Publishers, 1993.

[12] I. Stainvas and N. Intrator. Blurred face recognition via a hybrid network architecture. In *Proc. ICPR*, volume 2, pages 809–812, 2000.

[13] G. S. Berns, J. D. Cohen, and M. A. Mintun. Brain regions responsive to novelty in the absence of awareness. *Science*, 276:1272–1276, 1997.

[14] B. Landau, L. B. Smith, and S. Jones. The importance of shape in early lexical learning. *Cognitive Development*, 3:299–321, 1988.

[15] S. Becker and M. Plumbley. Unsupervised neural network learning procedures for feature extraction and classification. *Applied Intelligence*, 6:185–203, 1996.

[16] G. E. Hinton and T. J. Sejnowski, editors. *Unsupervised Learning: Foundations of Neural Computation*. MIT Press, Cambridge, MA, 1999.