Unsupervised learning of visual structure

Shimon Edelman[1], Nathan Intrator[2,3], and Judah S. Jacobson[4]

[1] Department of Psychology
Cornell University, Ithaca, NY 14853, USA
[2] Institute for Brain and Neural Systems
Brown University, Providence, RI 02912, USA
[3] School of Computer Science
Tel-Aviv University, Tel Aviv 69978, Israel
[4] Department of Mathematics
Harvard University, Cambridge, MA 02138, USA

Abstract. To learn a visual code in an unsupervised manner, one may attempt to capture those features of the stimulus set that would contribute significantly to a statistically efficient representation (as dictated, e.g., by the Minimum Description Length principle). Paradoxically, all the candidate features in this approach need to be known before statistics over them can be computed. This paradox may be circumvented by confining the repertoire of candidate features to actual scene fragments, which resemble the "what+where" receptive fields found in the ventral visual stream in primates. We describe a single-layer network that learns such fragments from unsegmented raw images of structured objects. The learning method combines fast imprinting in the feedforward stream with lateral interactions to achieve single-epoch unsupervised acquisition of spatially localized features that can support systematic treatment of structured objects [1].

1 A paradox and some ways of resolving it

It is logically impossible to form a principled structural description of a visual scene without prior knowledge of related scenes. Adapting an observation made by R. A. Fisher, such knowledge must, in the first instance, be statistical. Several recent studies indeed showed that subjects are capable of unsupervised acquisition of statistical regularities (e.g., conditional probabilities of constituents) that can support structural interpretation of novel scenes composed of a few simple objects [2, 3]. Theoretical understanding of unsupervised statistical learning is, however, hindered by a paradox perceived as "monstrous and unmeaning" already in the Socratic epistemology: statistics can only be computed over a set of candidate primitive descriptors if these are identified in advance, yet the identification of the candidates requires prior statistical data (cf. [4]).[1]

Figure 1 illustrates the paradox at hand in the context of scene interpretation. To decide whether the image on the left is better seen as containing horses (and riders) rather than centaurs requires tracking the representational utility of `horse` over a sequence of images. But for that one must have already acquired

the notion of `horse` — an undertaking that we aimed to alleviate in the first place, by running statistics over multiple stimuli. In what follows, we describe a way of breaking out of this vicious circle, suggested by computational and neurobiological considerations.

Fig. 1. An intuitive illustration of the fundamental problem of unsupervised discovery of the structural units best suited for describing a visual scene (cf. *Left*). Is the being in the forefront of this picture integral or composite? The visual system of the Native Americans, who in their first encounter reportedly perceived mounted Spaniards as centaur-like creatures (cf. [5], p.127), presumably acted on a principle that prescribes an integral interpretation, in the absence of evidence to the contrary. A sophisticated visual system should perceive such evidence in the appearance of certain candidate units in multiple contexts (cf. *Middle*, where the conquistadors are seen dismounted). Units should not have to appear in isolation (*Right*) to be seen as independent.

1.1 Computational considerations

The choice of primitives or features in terms of which composite objects and their structure are to be described is the central issue at the intersection of high-level vision and computational learning theory. Studies of unsupervised feature extraction (see e.g. [6] for a review) typically concentrate on the need for supporting recognition, that is, telling objects apart. Here, we are concerned with the complementary need — seeking to capture commonalities between objects — which stems from the coupled constraints of making explicit object structure, as per the principle of systematicity [1], and maintaining representational economy, as per the Minimum Description Length (MDL) principle [7].

One biologically relevant representational framework that aims for systematicity while observing parsimony is the Chorus of Fragments (CoF [8, 1]). In the CoF model, the graded responses of "what+where" cells [9, 10] coarsely tuned both to shape and to location form a distributed representation of stimulus structure. In this paper, we describe a method for unsupervised acquisition of "what+where" receptive fields from examples.

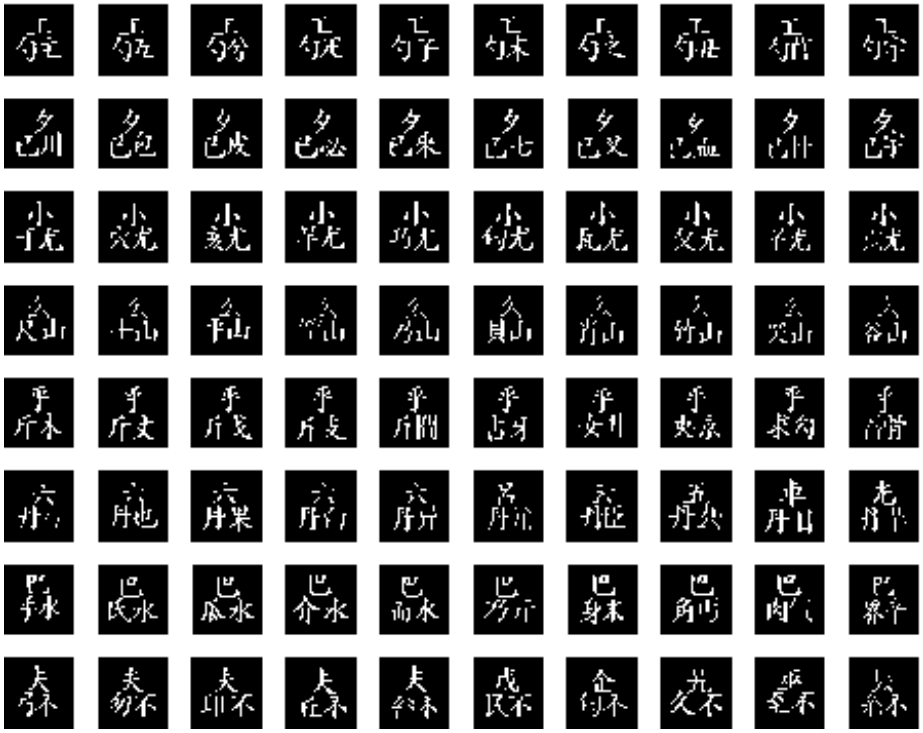

Fig. 2. The challenge of unsupervised learning of shape fragments that could be useful for representing structured objects is exemplified by this set of 80 images, showing triplets of Kanji characters. A recent psychophysical study [3] showed that observers unfamiliar with the Kanji script learn representations that capture the pair-wise conditional probability between the characters over this set, tending to treat frequently co-occurring characters as wholes. This learning takes place after a single exposure to the images in a randomly ordered sequence.

To appreciate the challenges inherent in the unsupervised structural learning task, consider the set of 80 images of triplets of Kanji characters appearing in Figure 2. A recent psychophysical study showed that observers unfamiliar with the Kanji script learn representations that capture subtle statistical dependencies among the characters, after being exposed to a randomly ordered sequence of these images just once [3]. When translated into a constraint on the functional architecture of the CoF model, this result calls for a fast *imprinting* of the feedforward connections leading to the "what+where" cells. Another requirement, that of *competition* among same-location cells, arises from the need to achieve a sufficient diversity of the local shape basis. Finally, *cooperation* among far-apart cells, seems to be necessary to detect co-occurrences among spatially distinct fragments.

The latter two requirements can be fulfilled by lateral connections [11] whose sign depends on the retinotopic separation between the cells they link. Although lateral connections play a central role in many approaches to feature extraction [6], their role is usually limited to the orthogonalization of the selectivities of different cells that receive the same input. In one version of our model, such short-range inhibition is supplemented by longer-range excitation, a combination that is found in certain models of low-level vision (see the review in [11]). These lateral connections are relevant, we believe, to the understanding of neural response properties and plasticity higher up in the ventral processing stream, in areas V4 and TEO/TE.

1.2 Biological motivation

We now briefly survey the biological support for the functional model proposed above.

- *Joint coding of shape and location information.* Cells with "what+where" receptive fields, originally found in the prefrontal cortex [9], are also very common in the inferotemporal areas [10].
- *Lateral interactions.* The anatomical substrate for the lateral interactions proposed here exists in the form of "intrinsic" connections at all levels of the visual cortical hierarchy [12]. Physiologically, the "inverted Mexican hat" spatial pattern of lateral inputs converging on a given cell, of the kind used in our first model (described in section 2.1) is consistent with the reports of selective connections linking V1 cells with like response properties (see, e.g., the chapter by Polat et al. in [11]). The specific role of neighborhood (lateral) competition in shaping the response profiles of TE neurons is supported by findings such as that of selective augmentation of neuron responses by locally blocking GABA, a neurotransmitter that mediates inhibition [18].
- *Fast learning.* Fast synaptic modification following various versions of the Hebb rule [13], which we used in one of the models described below, has been reported in the visual cortex [14] and elsewhere in the brain [15]. Evidence in support of the biological relevance of the other learning rule we tested, BCM [16] is also available [17].

2 Learning "what+where" receptive fields

Intuitively, spatial ("where") selectivity of the "what+where" cell can be provided by properly weighting its feedforward connections, so as to create a window corresponding to a fragment of the input image; shape ("what") selectivity can then be obtained by fast learning (ultimately from a single example) that would create, within that window, a template for the stimulus fragment. The networks we experimented with consisted of nine groups of such cells, arranged on a loose grid (Figure 3, left). In the experiments described here the networks contained either 3 or 8 cells per location. Each cell saw the entire input image through

a window corresponding to the cell's location; for reasons of biological plausibility, the windows were graded (Gaussian; see Figure 4, left). Results obtained with the two learning rules we studied, of the Hebbian and BCM varieties, are described in sections 2.1 and 2.2, respectively.

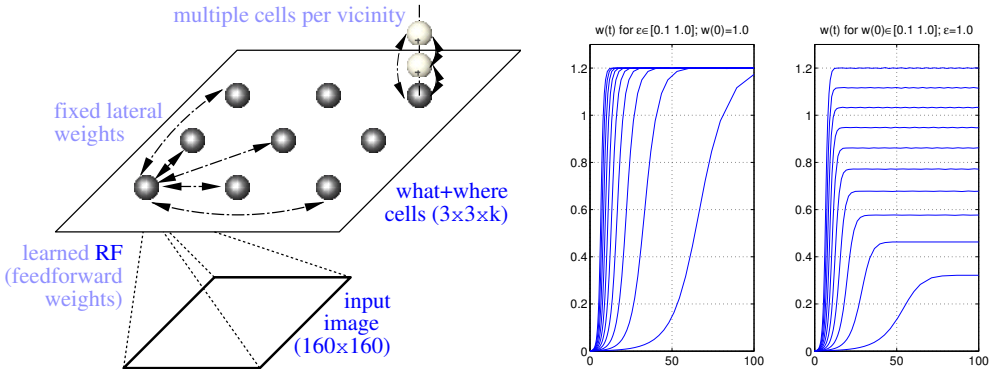

Fig. 3. *Left:* the Hebbian network consisted of groups of "what+where" cells arranged retinotopically, on a 3×3 loose grid, over the input image. Each cell received the 160×160 "retinal" input, initially weighted by a Gaussian window (Figure 4, left). In addition, the cells were fully interconnected by lateral links, weighted by a difference of Gaussians, so that weights between nearby cells were inhibitory, and between far-apart cells excitatory (Figure 4, right). *Right:* a numerical solution for the feedforward connection weight $w(t)$ given a constant input $x = 1$, with the learning rate ϵ (left pane) and $w(0)$ (right pane) varying in 10 steps between 0.1 and 1 (see eqns. 1 and 2).

2.1 Hebbian learning

For use with the Hebbian rule, the "what+where" cells were fully interconnected by lateral links with weights following the inverted Mexican hat profile (Figure 4, right). Given an input image $\mathbf{x}$, the output of cell i was computed as:

$$\begin{aligned} y_i &= \tanh(c_+) \\ c_+ &= c \ \mathrm{sign}(c) \\ c &= (\mathbf{x}.\mathbf{w}_i + \textstyle\sum_{j \neq i} v_{ij} y_j) - \theta \\ \theta(t) &= 0.5(\max\{c(t-h), \ldots, c(t-1)\} - \min\{c(t-h), \ldots, c(t-1)\}) \end{aligned} \quad (1)$$

where $\mathbf{w}_i$ is the synaptic weight vector, $\theta(t)$ is a history-dependent threshold (set to the mean of the last h values of c), $v_{ij} = G(d_{ij}, 1.6\sigma) - G(d_{ij}, \sigma)$ is the strength of the lateral connection between cells i and j; $G(x, \sigma)$ is the value at x of a Gaussian of width σ centered at 0 (the dependence of v on d is illustrated in Figure 4, right).

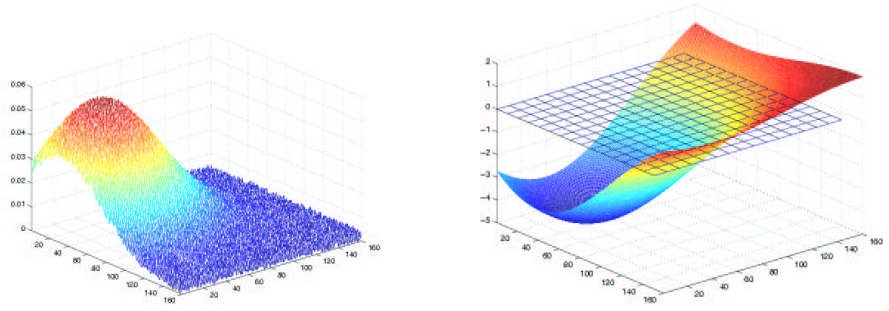

Fig. 4. *Left:* the initial (pre-exposure) feedforward weights constituting the receptive field (RF) of the cell in the lower left corner of the 3×3 grid (cf. Figure 3). The initial RF had the shape of a Gaussian whose standard deviation was 40 pixels (equal to the retinal separation of adjacent cells on the grid). The centers of the RFs were randomly shifted by up to ±10 pixels in each retinal coordinate according to a uniform distribution. The Gaussian was superimposed on random noise with amplitude uniformly distributed between 0 and 0.01. *Right:* the lateral weights in the Hebbian network, converging on the cell whose initial RF is shown on the left, plotted as a function of the retinotopic location of the source cell.

The training consisted of showing the images to the network in a random order, in a single pass (epoch), as in the psychophysical study [3]. Each input was held for a small number of "dwell cycles" (2-5), to allow the lateral interactions to take effect. In each such cycle, the feedforward weights w_{mn} for pixels x_{mn} were modified according to this rule:

$$w_{mn}(t+1) = w_{mn}(t) + \eta(yx_{mn}(t)w_{mn}(0) - y^2 w_{mn}(t)) \tag{2}$$

In this rule, the initial (Gaussian) weight matrix, $\mathbf{w}(0)$, determines the effective synaptic modification rate throughout the learning process. To visualize the dynamics of this process, we integrated eq. 2 numerically; the results, plotted in Figure 3, right, support the intuition just stated. Note that eq. 2 resembles Oja's self-regulating version of the Hebbian rule, and is local to each synapse, hence particularly appealing from the standpoint of biological modeling. Note also that the dynamic nature of the threshold $\theta(t)$ and the presence of a nonlinearity in eq. 1 resemble the BCM rule of [19].

The receptive fields (RFs) of the "what+where" cells acquired in a typical run through a single exposure of the Hebbian network to a randomly ordered sequence of the 80 images of Figure 2, are shown in Figure 5. Characters more frequent in the training set (such as the ones appearing in the top locations in Figure 2) were generally the first to be learned. Importantly, the learned RFs are relatively "crisp," with the template for one (or two) of the characters from the training data standing out clearly from the background. Pixels from other characters are attenuated (and can probably be discarded by thresholding). A parametric exploration determined that (1) the learning rate η in eq. 2 had to

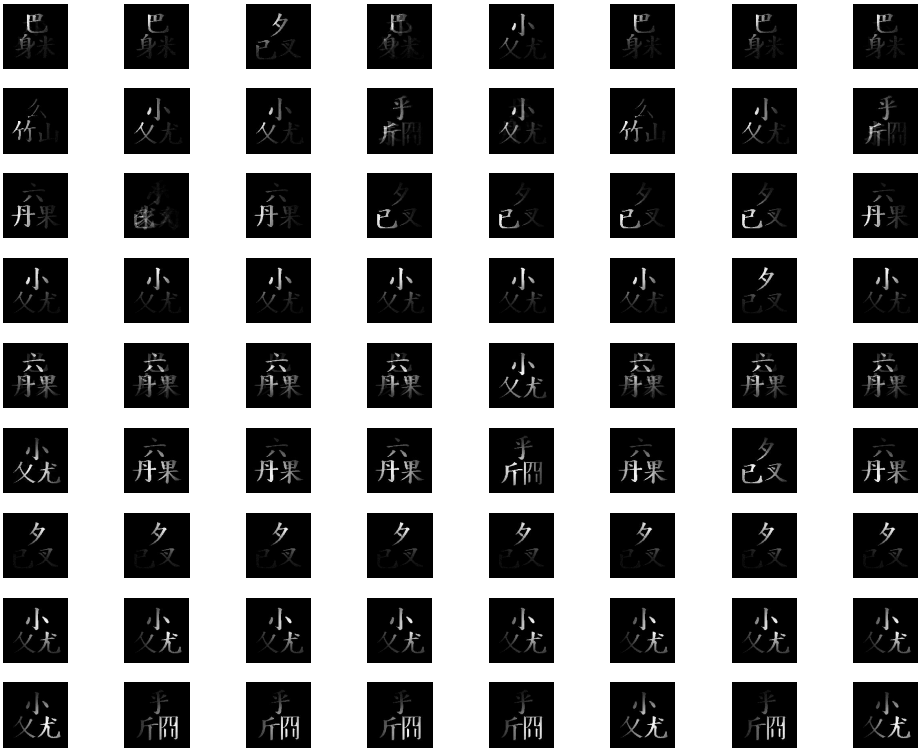

Fig. 5. The receptive fields of a 72-cell Hebbian network (8 cells per location) that has been exposed to the images of Figure 2. Each row shows the RFs formed for one of the image locations.

be close to 1.0 for meaningful fragments to be learned; (2) the results were only slightly affected by varying the number of dwell cycles between 2 and 20; (3) the formation of distinct RFs for the same location depended on the competitive influence of the lateral connections.

To visualize concisely the outcome of 20 learning runs of the network (equivalent to running an ensemble of 20 networks in parallel), we submitted the resulting 1440 RFs (each of dimensionality $160 \times 160 = 25600$) to a k-means routine, set to extract 72 clusters. Among the RFs thus identified (Figure 6), one finds templates for single-character shapes (e.g., #1, 14), for character pairs with a high mutual conditional probability in the data set (e.g., #7), an occasional "snapshot" of an entire triplet (#52), as well as a minority of RFs that look like a mix of pixels from different characters (#50, 51). Note that even these latter RFs could serve as useful features for projecting the data on, given the extremely high dimensionality of the raw data space (25600).

The MDL and related principles [20, 7] suggest that features that tend to cooccur frequently should be coded together. To assess the sensitivity of our RF

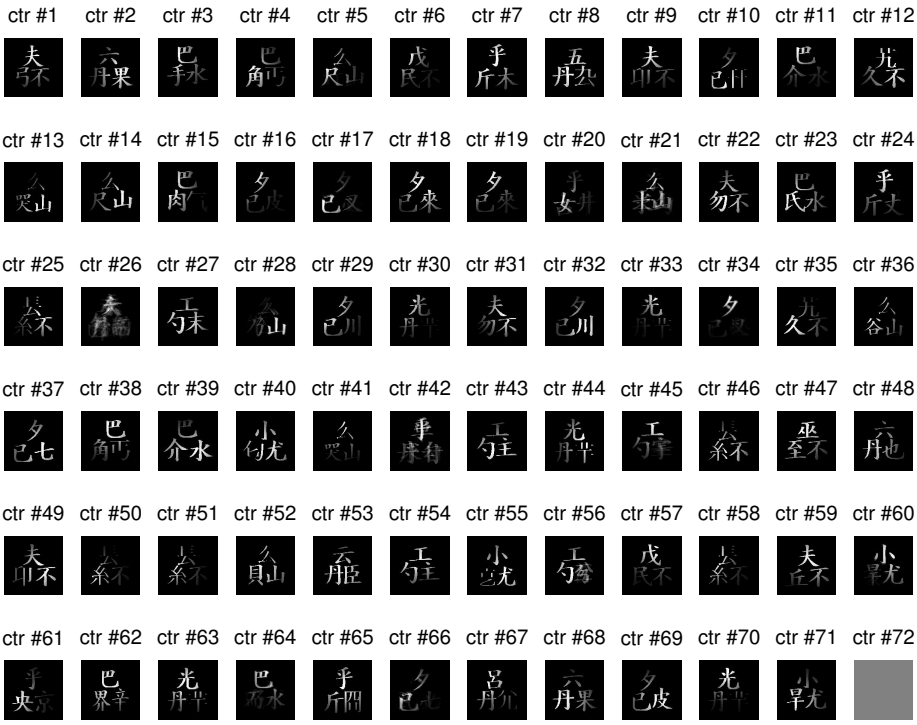

Fig. 6. The 72 RFs that are the cluster centroids identified by a k-means procedure in a set of 1440 RFs (generated by 20 runs of a $3 \times 3 \times 8$ Hebbian network, exposed to the images of Figure 2. See text for discussion.

learning method to the statistical structure of the stimulus set, we calculated the number of RFs acquired in the 20 learning runs, for each of the four kinds of input patterns whose occurrences in Figure 2 were controlled (for the purposes of an earlier psychophysical study [3]). The patterns could be of "fragment" or "composite" kind (consisting of one or two characters, respectively), and could belong to a pair that appeared together always (conditional probability of 1) or in half of the instances ($CP = 0.5$). The RF numbers reflected these probabilities, indicating that the algorithm was indeed sensitive to the statistics of the data set.

To demonstrate that the learning method developed here can be used with gray-level images of 3D objects (and not only with binary character images), we ran a 27-unit network (3 cells per location) on the 36 images of composite shapes shown in Figure 7, top. As with the character images, the network proved capable of extracting fragments corresponding to meaningful parts (Figure 7, bottom; e.g., #1, 19) or to combination of such parts (e.g., #4, 13).

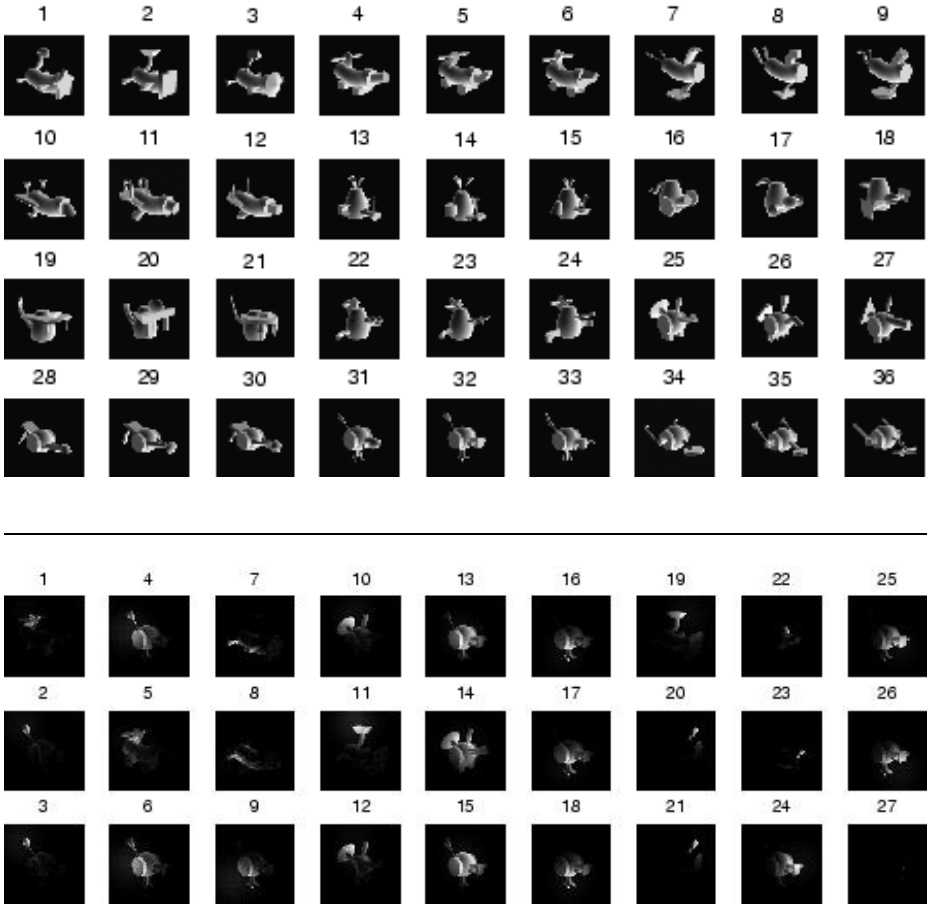

Fig. 7. Fribbles. *Top:* 36 images of "fribbles" (composite objects available for download from http://www.cog.brown.edu/∼tarr/stimuli.html. *Bottom:* fragments extracted from these images by a $3 \times 3 \times 3$ Hebbian network of "what+where" cells.

2.2 BCM learning

The second version of the model learned its RFs by optimizing a BCM objective function [16] using simple batch-mode gradient descent with momentum. The total gradient was computed as a weighted sum of the gradient contributions from the feedforward BCM learning rule, a lateral inhibition term, and the norm of the weights. The lateral inhibition pattern was uniform: activations were inhibited by a constant sum of the activations of the other neurons. A sigmoidal transfer function (tanh) was then applied to this modified activation in order to prevent individual activations from growing too high. The limiting value of this nonlinearity controls the minimal probability of the event to which a neuron

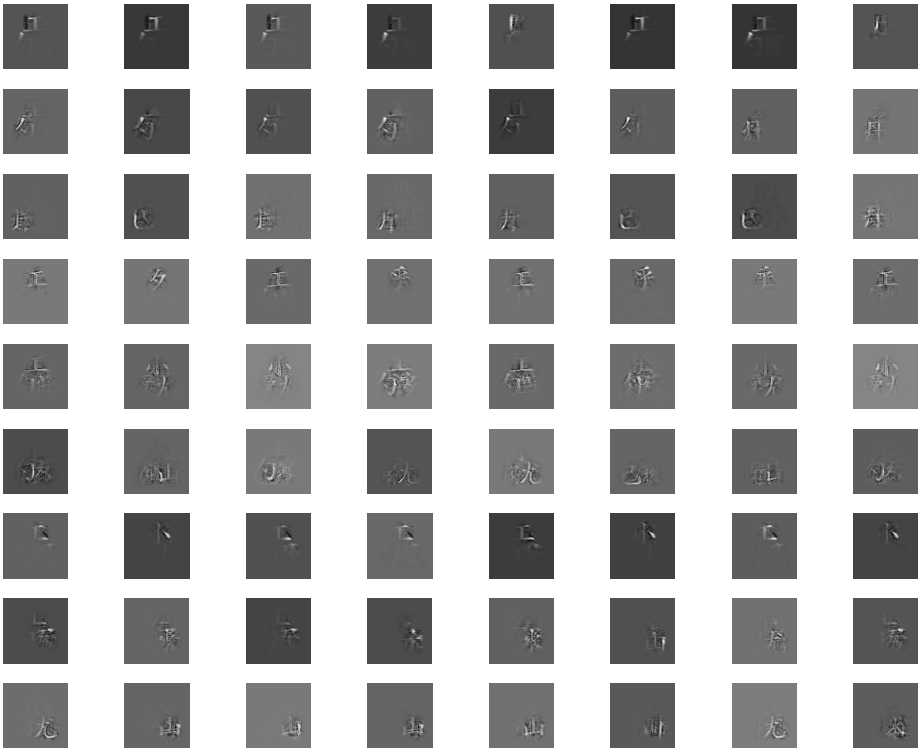

Fig. 8. The receptive fields of a 72-cell BCM network (8 cells per location) that has been exposed to the Kanji character images of Figure 2.

becomes selectively tuned [16]. By limiting the activation to about 10, events with probability of about 1/10 could be found (without this step, each neuron would eventually converge to one of the individual inputs).

The activity of neuron k in the BCM network is $c_k = \mathbf{x}.\mathbf{w}_k$, where $\mathbf{w}_k$ is its synaptic weight vector. The inhibited activity of the k'th neuron and its threshold are:

$$\tilde{c}_k = c_k - \eta \sum_{j \neq k} c_j \qquad \tilde{\Theta}_M^k = E[\tilde{c}_k^2] \tag{3}$$

where $E[\cdot]$ denotes expectation. When the nonlinearity of the neuron is included, the inhibited activity is given by: $\tilde{c}_k = \tanh(c_k - \eta \sum_{j \neq k} c_j)$, and the learning rule becomes:

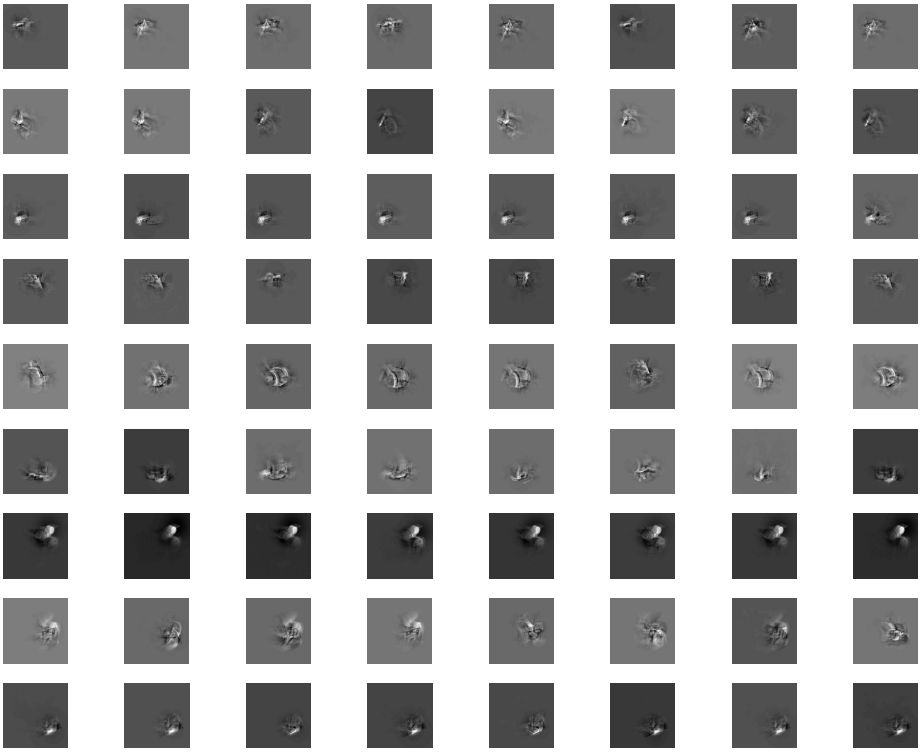

Fig. 9. The receptive fields of a 72-cell BCM network (8 cells per location) that has been exposed to the fribble images of Figure 7.

$$\dot{\mathbf{w}}_k = \mu \left(E\left[\phi(\tilde{c}_k, \tilde{\Theta}_M^k)\sigma'(\tilde{c}_k)x\right] - \eta \sum_{j \neq k} E\left[\phi(\tilde{c}_j, \tilde{\Theta}_m^j)\sigma'(\tilde{c}_j)x\right] \right) - \lambda \|\mathbf{w}_k\| \quad (4)$$

where σ' is the derivative of tanh, $\phi(c, \theta) \doteq c(c - \theta)$ [16]; μ and η are learning rates, the last term is the weight decay, and λ is a small regularization parameter determined empirically. Note that the lateral inhibition network performs a search of k-dimensional projections jointly; thus may find a richer structure, which a stepwise approach might miss [21].

The RFs learned by a 72-cell ($3 \times 3 \times 8$) BCM network are shown in Figures 8 and 9. As with the Hebbian network, individual characters and fribble fragments were picked up and imprinted onto the RFs of the neurons.

To compare the performance of the two biologically motivated statistical learning methods to a well-known benchmark, we carried out an independent component analysis (ICA) on the fribble image set, asking for 27 components.

Although it is not clear a priori that the existing ICA algorithms are suitable for our extremely high-dimensional learning problem, pixels belonging to distinct parts of the fribble objects *are* statistically independent and should be amenable to detection by ICA. The first 9 of the components extracted by a publicly available implementation of ICA are shown in Figure 10. By and large, these are not nearly as localized as the components learned by the Hebbian and the BCM methods, suggesting that their use as structural primitives would be limited. A full, quantitative investigation of the utility of distributed representations employing Hebbian, BCM and ICA features is beyond the scope of the present study.

Fig. 10. The first 9 of 27 independent components extracted from the fribbles image set (36 vectors of dimensionality 25600) by FastICA (http://www.cis.hut.fi/projects/ica/fastica/, courtesy of A. Hyvärinen); we used symmetrical decorrelation, 100 iterations, and the default settings for the other parameters.

3 Discussion

The unsupervised acquisition of meaningful shape fragments from raw, unsegmented image sequences exhibited by our networks is made possible by two of their properties: (1) fast feedforward learning, and (2) lateral interactions. In the Hebbian case, these characteristics are crucial: learning must be fast (it occurs within a single epoch, or, if the learning constant is too low, not at all), and the lateral interactions must combine local competition (to keep the representation sparse) with global cooperation (to capture sufficiently large chunks of objects). A parallel can be drawn between our space-variable lateral/Hebb rule and the use of lateral inhibition for feature decorrelation in unsupervised learning in general (e.g., [22, 19, 16]). The "lateral" interactions in algorithms such as the extended BCM [16] are not normally described in spatial terms. Interestingly, experience with our BCM implementation indeed suggests that lateral interactions incorporated into it need not be spatially variant to ensure useful behavior. An inquiry into the role of these parameters in learning spatial structure across multiple scales is currently under way in our lab.

Our models learn to find structure in raw images residing in a very high-dimensional space, which makes the problem extremely difficult [19]; yet, presenting the images in register with each other obviates the need to tolerate

translation, making the task much easier. In a more realistic setting, the learning would occur over base representations that are both more stable under stimulus transformations such as translation, and have lower dimensionality than the raw images. A biologically plausible modification to our models along these lines would involve feeding them the output of a simulated primary visual cortex, including simple, complex and hypercomplex cells, and employing space-variant resolution. Other challenges for the future include deriving our Hebbian learning rule from an objective function formulated from first principles such as MDL, and making its lateral interactions more realistic, e.g., by letting the network learn the strength of the lateral connections, perhaps using the same Hebbian mechanism as in the feedforward pathway. In the meanwhile, our results show that the paradox of unsupervised statistical learning can be circumvented: meaningful fragments of visual structure can be picked up from raw input by a joint application of computational and biological principles.

Notes

[1]The sense of paradox is well captured by the following passage from Plato's *Theaetetus* (360BC), in which Socrates points out the circularity in treating syllables as combinations of letters, if the latter are to be defined merely as parts of syllables:

Soc. ... there is one point in what has been said which does not quite satisfy me.
The. What was it?
Soc. What might seem to be the most ingenious notion of all: - that the elements or letters are unknown, but the combination or syllables known [...] can he be ignorant of either singly and yet know both together?
The. Such a supposition, Socrates, is monstrous and unmeaning.
Soc. But if he cannot know both without knowing each, then if he is ever to know the syllable, he must know the letters first; and thus the fine theory [that there can be no knowledge apart from definition and true opinion] has again taken wings and departed.
The. Yes, with wonderful celerity.

References

1. Edelman, S., Intrator, N.: Towards structural systematicity in distributed, statically bound visual representations. **-** (2001) – under review.
2. Fiser, J., Aslin, R.N.: Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychological Science **6** (2001) 499–504
3. Edelman, S., Hiles, B.P., Yang, H., Intrator, N.: Probabilistic principles in unsupervised learning of visual structure: human data and a model. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press (2002)
4. Gardner-Medwin, A.R., Barlow, H.B.: The limits of counting accuracy in distributed neural representations. Neural Computation **13** (2001) 477–504
5. Eco, U.: Kant and the Platypus. Secker & Warburg, London (1999)
6. Becker, S., Plumbley, M.: Unsupervised neural network learning procedures for feature extraction and classification. Applied Intelligence **6** (1996) 185–203

7. Bienenstock, E., Geman, S., Potter, D.: Compositionality, MDL priors, and object recognition. In Mozer, M.C., Jordan, M.I., Petsche, T., eds.: Neural Information Processing Systems. Volume 9. MIT Press (1997)
8. Edelman, S., Intrator, N.: (Coarse Coding of Shape Fragments) + (Retinotopy) ≈ Representation of Structure. Spatial Vision **13** (2000) 255–264
9. Rao, S.C., Rainer, G., Miller, E.K.: Integration of what and where in the primate prefrontal cortex. Science **276** (1997) 821–824
10. Op de Beeck, H., Vogels, R.: Spatial sensitivity of Macaque inferior temporal neurons. J. Comparative Neurology **426** (2000) 505–518
11. Sirosh, J., Miikkulainen, R., Choe, Y., eds.: Lateral Interactions in the Cortex: Structure and Function. electronic book (1995) http://www.cs.utexas.edu/users/nn/lateral_interactions_book/cover.html.
12. Lund, J.S., Yoshita, S., Levitt, J.B.: Comparison of intrinsic connections in different areas of macaque cerebral cortex. Cerebral Cortex **3** (1993) 148–162
13. Brown, T.H., Kairiss, E.W., Keenan, C.L.: Hebbian synapses: biophysical mechanisms and algorithms. Ann. Rev. Neurosci. **13** (1990) 475–511
14. Fregnac, Y., Schulz, D., Thorpe, S., Bienenstock, E.: A cellular analogue of visual cortical plasticity. Nature **333** (1988) 367–370
15. Gluck, M.A., Granger, R.: Computational models of the neural bases of learning and memory. Ann. Rev. Neurosci. **16** (1993) 667–706
16. Intrator, N., Cooper, L.N.: Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. Neural Networks **5** (1992) 3–17
17. Bear, M.F., Malenka, R.C.: Synaptic plasticity: LTP and LTD. Curr. Opin. Neurobiol. **4** (1994) 389–399
18. Wang, Y., Fujita, I., Murayama, Y.: Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. Nature Neuroscience **3** (2000) 807–813
19. Intrator, N.: Feature extraction using an unsupervised neural network. Neural Computation **4** (1992) 98–107
20. Barlow, H.B.: Unsupervised learning. Neural Computation **1** (1989) 295–311
21. Huber, P.J.: Projection pursuit (with discussion). The Annals of Statistics **13** (1985) 435–475
22. Földiák, P.: Learning invariance from transformation sequences. Neural Computation **3** (1991) 194–200