Towards structural systematicity in distributed, statically bound visual representations

Shimon Edelman
Department of Psychology
232 Uris Hall, Cornell University
Ithaca, NY 14853-7601, USA

Nathan Intrator
Institute for Brain and Neural Systems
Box 1843, Brown University
Providence, RI 02912, USA*

July 2001
revised March 2002, October 2002

Abstract

The problem of representing the spatial structure of images, which arises in visual object processing, is commonly described using terminology borrowed from propositional theories of cognition, notably, the concept of compositionality. The classical propositional stance mandates representations composed of symbols, which stand for atomic or composite entities and enter into arbitrarily nested relationships. We argue that the main desiderata of a representational system — productivity and systematicity — can (indeed, for a number of reasons, should) be achieved without recourse to the classical, proposition-like compositionality. We show how this can be done, by describing a systematic and productive model of the representation of visual structure, which relies on static rather than dynamic binding and uses coarsely coded rather than atomic shape primitives.

1 Problematic issues in the representation of structure

The focus of theoretical discussion in visual object processing, which for a long time concentrated on recognition-related problems such as viewpoint invariance, has recently started to shift to the representation of object structure, following a particularly forceful statement of the issue by (Hummel, 2000). View- or appearance-based solutions to recognition-related problems developed and tested on a variety of object classes (Ullman, 1998) have been termed "holistic" because they do not explicitly represent or act upon the internal relational structure of objects (Edelman, 1998). As such, view-based recognition models are expected to be unable to deal with issues of visual structure. Borrowing an example from (Hummel, 2000), a holistic system can be made to discriminate an object composed of a circle atop a square from another one in which the square is on top of the circle, but it is not likely to be able to report that these two objects are also structurally similar (i.e., that they share parts).

This argument can be related to broader issues in cognitive science by invoking the observation that the main difficulties in the processing of structure lie in achieving productivity and, more importantly, systematicity, two traits commonly attributed to human cognition in general (Fodor and Pylyshyn, 1988; Fodor,

*On leave from the School of Computer Science, Tel-Aviv University, Tel Aviv 69678, Israel.

1998). As it happens, the best-known theory of structure representation in vision (Biederman, 1987; Hummel and Biederman, 1992) is both productive and systematic — as are all *compositional* approaches that construe objects as composed of a small number of generic parts conjoined by universally applicable categorical relations. Nevertheless, we consider the issue of systematicity in vision to be far from settled, in view of the assumptions made by the classical compositional approaches.

The most crucial of these is the assumption of the possibility of dynamic binding of parts to slots in relational structures. Any model that incorporates this assumption is equivalent in its computational power to λ-calculus, and therefore to a Turing Machine (Church, 1941). Considerations of parsimony dictate that the computational mechanism behind a model be not more powerful than absolutely necessary, lest the modeling be reduced to an exercise in programming. In the context of biological vision, considerations that discourage appeals to general-purpose computation or symbol manipulation are especially pertinent, given the controversial status of neuronal-level theories of dynamic binding (Roskies, 1999). Consequently, we chose *not* to assume that dynamic binding can be used to model the processing of structure in primate vision, or, by implication, that the problem of systematicity in vision has been already solved by the classical theories.

In this paper, we examine the possibility of achieving systematicity within an approach to structure representation that relies on static binding and is based on an existing theory of recognition and categorization (Edelman, 1999). We first describe an implemented computational model that uses distributed representations and static binding to attain a degree of systematicity on unfamiliar stimuli. We then survey data from the psychology and the neurobiology of visual structure representation suggesting that the systematicity in biological vision is limited in a manner implied by the proposed computational framework. A formal definition of visual systematicity and a discussion of its relationships to computational semantics are offered in an appendix.

1.1 The problem of productivity

Intuitively, a cognitive system is productive if it is open-ended, that is, if the set of entities with which it can deal is, at least potentially, infinite. The most often-cited example of a productive system is language: people both produce and comprehend quite an impressive variety of sentences, with no clear limit to their number. In the context of visual recognition, the productivity challenge is to represent an unlimited variety of objects in a finite-resource system.

The stipulation that the representational system rely on finite resources rules out any solution to the productivity problem that is based on rote memory. If, however, the system can interpolate among examples stored in memory, its representational capabilities can be greatly extended (Poggio and Hurlbert, 1994). For instance, novel views of familiar objects can be recognized by interpolation among a few familiar views (Poggio and Edelman, 1990), and novel instances of familiar shape classes can be categorized by interpolation among stored class prototypes (Edelman and Duvdevani-Bar, 1997a). Moderately novel categories too can be dealt with, on the basis of analogy to familiar ones (Edelman and Duvdevani-Bar, 1997b).

Within this framework, radically novel categories can only be treated by extending the system's repertoire of stored prototypes. Although it may seem that the need for extra memory would make the achievement of productivity problematic, this does not have to be so. Postulating a *fixed* upper limit on the resources to be used unduly constrains the range of possible approaches to productivity. Indeed, a significantly sublinear (e.g., logarithmic) growth of the required resources with the size of the problem can usually be accommodated in practice, as discussed in any textbook on computational complexity (Aho et al., 1974).

In view of this abstract (computational-level) observation, and given the existing practical approaches to

recognition and categorization that exhibit productivity, we consider this problem to be of lesser importance than systematicity, which is the focus of this paper.

1.2 The problem of systematicity and its relationship to classical compositionality

In the general context of cognition, the problem of systematicity has been described by (Fodor and McLaughlin, 1990), p.184, thus:

> [A]s a matter of psychological law, an organism is able to be in one of the states belonging to the family only if it is able to be in many of the others [...] You don't find organisms that can think the thought that the girl loves John but can't think the thought that John loves the girl.

Systematicity is commonly considered to be the crux of the debate focusing on the representational theory of mind (Fodor, 1987; Kaye, 1995). Despite the importance of this concept, the debate surrounding it has probably been less productive (and less systematic) than it could be, because systematicity "has rarely been well defined" (Prinz, 1994). The following definition seems to be the closest the field has come to a formal approach:

Definition 1 (Hadley, 1997, p.140) *A cognitive agent, C, exhibits systematicity just in case its cognitive architecture causally ensures that C has the capacity to be in a propositional attitude (A) towards proposition aRb if and only if C has the capacity to be in attitude (A) towards proposition bRa.*

The systematicity of human verbal behavior with respect to simple sentences such as "John loves Mary" is considered by many to be a proof that cognition is *compositional*. In its classical mathematical formulation (Frege, 1993), the principle of compositionality is the idea that the meaning of a whole (say, a string of symbols) is a function of the meaning of its parts (i.e., of the individual symbols).[1] In non-technical writing in cognitive science, compositionality came to be construed as a recipe for building complex structures out of simpler ones, by concatenation-like operations. More often than not, it is considered as a challenge to be met by a sophisticated representational framework, along with productivity and systematicity (Fodor and Pylyshyn, 1988; Bienenstock, 1996; Bienenstock et al., 1997). In particular, Fodor and Pylyshyn (1988, p.41) tend not even to distinguish between compositionality and systematicity, remarking that "perhaps they should be viewed as aspects of a single phenomenon."

We contend that this approach constitutes a category mistake. Generally speaking, compositionality does not belong in the same class of phenomena as systematicity, because it is not a problem in the same sense that productivity and systematicity are. A classical (Fregean) compositional representation, in which symbols standing for a few generic and primitive (atomic) entities enter into arbitrarily nested relations, would be automatically both productive and systematic (just as propositional calculus is). One must realize, therefore, that compositionality is a (possible) means, not an end in itself: if productivity and systematicity can be attained in some other fashion, the need for classical compositionality would be obviated. Thus, the standard "proof" that assumes systematicity and infers compositionality, which Fodor repeatedly describes as an "argument from the best explanation," commits a logical fallacy (*petitio principii*), akin to claiming that birds must be lighter than air because they can fly; cf. (van Gelder, 1990), p.378.

In vision, the challenge of systematicity inherent in Definition 1 is typically illustrated by examples such as those in Figure 1: a system that can make sense of object $A = (\texttt{circle}\ above\ \texttt{square})$ should also be able to make sense of $B = (\texttt{square}\ above\ \texttt{circle})$ (Hummel, 2000). This propositional construal

[1] A discussion of formal compositionality in the context of computational semantics appears in appendix A.

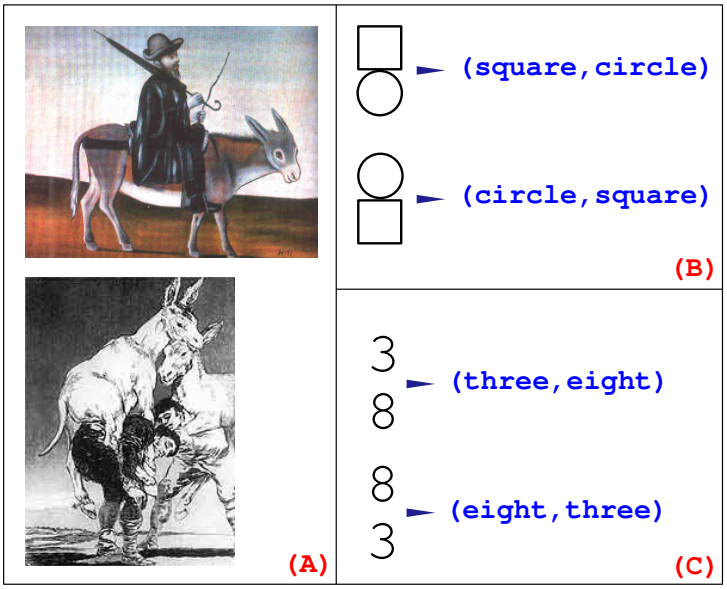

Figure 1: According to a classical definition, a representation is systematic if its ability to encode some combination of symbolic constituents entails the ability to encode other combinations of the same constituents (section 1.2). This figure illustrates a straightforward extension of such combinatorial systematicity to vision. (A): a system that can represent a man riding a donkey should also be able to represent a donkey riding a man (top: *A Physician Riding a Donkey*, by Niko Pirosmanashvili; bottom: *You Who Can't Do Anything*, by Francisco Goya). (B): the call for systematicity implicit in (Hummel, 2000) uses the example depicted here. (C): the computational experiments reported below use composites consisting of pairs of numeral shapes to test for systematicity.

of systematicity, which relies ultimately on concepts imported from symbolic logic, is problematic, for two reasons.

First, the very coaching of the problem of representation in propositional terms amounts to adopting the ontological stance of reifying discrete parts (cf. Figure 2), which in Definition 1 are designated by symbols a and b. This stance may, in fact, not be defensible in vision (Edelman, 2002). Because of the pixellated nature of the lowest-level representation in any visual system that involves a spatial sampling of images, any visual object can be thought of, trivially, as composed of discrete "parts." Such parts, however, need not be in any sense real, and need not afford any useful insight into the object's nature (Smith, 2001). Moreover, (images of) many objects — even those designated in English by count nouns, such as boots, shellfish, apples, worms, branches — are not readily decomposable into middle-scale parts (as opposed to pixels).

Second, even for objects that are clearly decomposable into middle-scale parts, the need for representing these parts rearranged according to a different plan is questionable. For example, a face is typically perceived as a particular arrangement of two eyes, a nose and a mouth; the need to represent an object composed of these same parts in a different configuration virtually never arises outside the laboratory. Likewise, an image of a quadruped animal is normally seen to possess a head, a body, and four legs; a random rearrangement of these would, however, no longer count as an image of an animal.

Despite these problems, situations clearly exist in which the ability to represent middle-scale parts (in a member of category for which such a description is appropriate) is important and must be insisted upon.

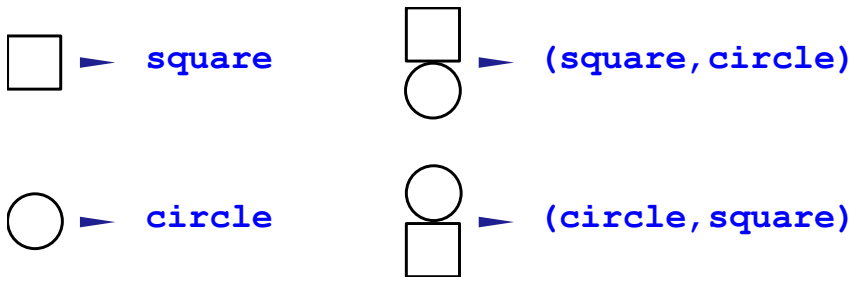

Figure 2: In vision, structural descriptions (a compositional approach, in which object shapes are described in terms of relatively few generic components, joined by categorical spatial relationships chosen from an equally small fixed set) is widely regarded as the only computational theory that is at all capable of delivering productivity and systematicity (Biederman, 1987; Bienenstock et al., 1997; Hummel, 2000). This illustration shows a hypothetical solution to the systematicity problem based on the principle of compositionality. Let us denote a square shape by A and a circle by B. Without loss of generality (cf. (Zadrozny, 1994), section 2), we may only treat here the relation *above*, which we denote by the symbol $\natural$ (e.g., $A\natural B$ standing for "A above B"). *Left:* an interpretation function, f, maps the two primitive shapes into their corresponding atomic symbols, as follows: $f(A) = a$, $f(B) = b$. *Right:* the systematicity of the classical compositional scheme is manifested in its ability to assign an intuitively correct interpretation to $B\natural A \xrightarrow{f} b.a$, given that it can interpret another shape, $A\natural B \xrightarrow{f} a.b$ consisting of the same components and utilizing the same relation ("above"). The dot in $a.b$ denotes concatenation. See section 1.2 for a discussion.

For example, for shape classes such as quadruped animals, the detection of the common parts (head, legs) may aid superordinate-level categorization. Typically, in such cases the parts are specific to a superordinate category (rather than generically applicable across all categories), and are likely to be functionally significant.

The constraint on representation implied by these latter observations may be termed *context systematicity*: a system that has the notion of a *head* (a part of an animal – say, a cow) or a *handset* (a part of a telephone), must also be able to detect the head of another animal (say, a pig), and the handset of an answering machine. We observe that this variety of systematicity is actually subsumed under Definition 1: if the relation aRb is taken to be asserting $head\_of(a, b)$ in the context of $head(a) \wedge cow(b)$, then bRa would be well-formed (albeit rather meaningless).

Lastly, the general-purpose *configurational systematicity* that is made explicit by Definition 1 is certainly applicable to the representation of *scenes* composed of discrete middle-scale objects. In this case, a straightforward translation of the definition into the visual domain is entirely unproblematic: a system that can represent the presence of a chair to the left of a table (aRb) in one scene must be capable of representing a table to the left of a chair (bRa) in another.

In the balance, despite serious conceptual problems surrounding it, visual systematicity rooted in Definition 1 has a broad intuitive appeal, which needs to be analyzed and acted upon. The action plan we develop in the rest of this paper has three components. First, we propose an intuitive formulation of compositionality tailored to vision and illustrate it by introducing a computational model of structure processing that employs distributed representations to address the core problem besetting classical, part-based systematicity – the nature of the parts in question. Second, we present a characterization of the limited systematicity exhibited by human observers gleaned from behavioral literature, along with neurobiological findings concerning the mechanisms that may support quasi-systematic visual perception. Third, we use the model, and the the-

ory of structure representation behind it, to derive concrete testable predictions for further behavioral and physiological experiments.

2 A computational scheme for distributed representation of structure

According to Definition 1, systematic treatment of a structured stimulus is ensured if any relevant relation defined on the representations of its constituents is equally applicable to all possible argument combinations: if aRb is well-defined, so must be bRa.[2] As we argued above, the visual world does not come equipped with *a priori* structuring tools such as the two-place relation R from the propositional calculus example. Representational systems typically *impose* structure on the visual world, and the main issue with which we are concerned here is how to do that within a philosophically reasonable and computationally feasible ontological framework that also exhibits at least some degree of systematicity. In this matter, theories such as Recognition By Components (Biederman, 1987; Hummel and Biederman, 1992) follow the example of propositional calculus, by adopting symbols standing for generic parts ("constituents") of objects, along with "crisp" categorical spatial relations such as *above* or *alongside*, as representational primitives. The resulting scheme is classically compositional (just as propositional calculus is) and therefore systematic, but its systematicity comes at a price, which we find too high, for reasons already discussed.

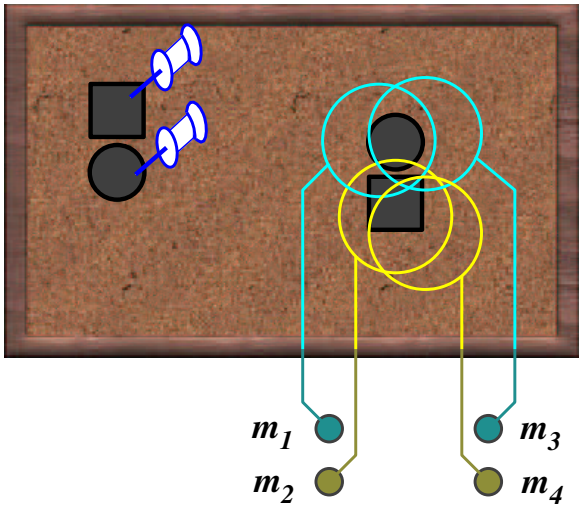

Figure 3: The Chorus of Fragments (CoF) model of structure representation (Edelman and Intrator, 2000) relies on the principles of distributed representation and of (static) binding by retinotopy. The circle and the square are bound to each other by virtue of appearing in those locations in the visual field (the "corkboard") that correspond to the receptive fields of the measurement units.

2.1 Coarse coding and binding by retinotopy: a Chorus of Fragments

An alternative, non-classical approach to systematicity is to represent an object by a set of non-categorical, non-generic measurements that are somehow spatially structured. We propose to use location in the visual

[2]Note that this propositional formulation does not require, say, that bRa be *true* if aRb is, only they both be *well-defined*. This point may be illustrated by letting R stand for the relation $>$ (*greater than*): the two propositions, $3 > 2$ and $2 > 3$, are both well-defined, albeit only one of them is true.

field in lieu of the abstract frame that encodes object structure. Intuitively, the constituents of an object are then bound to each other by virtue of residing in their proper places in the visual field; cf. (Clark, 2000), p.74. The visual field can then be thought of as a corkboard, whose spatial structure supports the arrangement of shape fragments pinned to it (Figure 3).[3] This scheme exhibits effective systematicity by virtue of its ability to represent different arrangements of the same constituents, as required by Definition 1. Coarse coding the constituents (e.g., representing each object fragment in terms of its similarities to some basis shapes) renders the scheme productive. Coarse coding also alleviates the ontological concerns that arise in classical, propositional schemes that postulate an alphabet of generic all-or-none parts. We call this approach to the representation of structure the Chorus of Fragments, or CoF (Edelman, 1999).

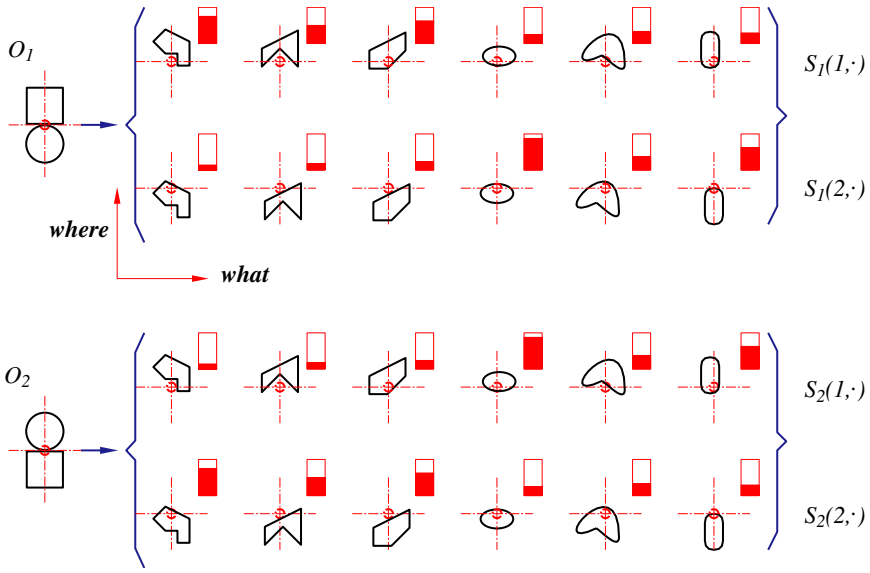

Figure 4: a Chorus of Fragments: encoding object structure by an ensemble of *what+where* cells (cf. section 4). Each such cell can be seen as performing a measurement m_i^l of the similarity between the stimulus and reference shape i at location l. In this illustration, six reference-shape measurements ($i \in \{1,2,3,4,5,6\}$) are replicated across two locations ($l \in \{1,2\}$), resulting in a 12-dimensional *what+where* representation. *Top:* an object $O_1 = A\natural B$, shown here as composed of fragments A and B, is mapped by the measurement functions $\{m_i^l\}$ into a 2×6 matrix S_1, whose rows correspond to measurements taken at different image locations. The values of $S_1(i,l)$ are depicted as the little "thermometers" associated with each reference shape. *Bottom:* the representation of object $O_2 = B\natural A$, which is structurally related to O_1. Configurational systematicity of this scheme with respect to objects $O_1 = A\natural B$ and $O_2 = B\natural A$ manifests itself in pairwise correlations between $S_1(1,\cdot)$ and $S_2(2,\cdot)$, and between $S_1(2,\cdot)$ and $S_2(1,\cdot)$, as discussed in section 2.4.

[3] Cf. Wittgenstein: "The essential nature of the propositional sign becomes very clean when we imagine it made up of spatial objects (such as tables, chairs, books) instead of written signs. The mutual spatial position of these things then expresses the sense of the proposition" (*Tractatus Logico-Philosophicus*, proposition 3.1431).

2.2 The principles behind the model

We now discuss in detail the constraints on the measurement system implied by the intuitive approach just presented (a formal treatment is given in the appendix). Consider the application of the proposed method to the square/circle systematicity challenge, illustrated schematically in Figure 4. The illustration shows the two "structured" objects from Hummel's example, represented by the outputs of six classes of measurement functions. Each class is parameterized by the location of the function's spatial support (in neurobiological terminology, its *receptive field*) with respect to the center of the visual field (marked by the cross-hairs).

Two instances of each class, corresponding to two locations, are shown; for example, the first measurement function in the top row is "tuned" to the same shape as the first one in the bottom row, but the two differ in the optimal location of the input. Effectively, each instance of a measurement function can be thought of as possessing a tuning curve both in the space of different possible objects (the *shape space*), and in the space of their locations (the "space" space). By virtue of these properties, the measurement functions bear an operational resemblance to the receptive fields of neurons, found at various stages of the mammalian visual system, that are tuned to stimulus features such as shape and location (Logothetis and Sheinberg, 1996; Tanaka, 1996); see section 4.

The central characteristics of this scheme of object representation are as follows:

C1 The choice of the *object fragments* to which the measurement functions are tuned should be governed by probabilistic principles such as Minimum Description Length (MDL), which dictates that the fragments support an informationally parsimonious coding of the data; see, e.g., (Zemel and Hinton, 1995). These principles are outside the scope of the present paper; for some initial psychophysical and computational explorations of this issue, see (Edelman et al., 2002; Edelman et al., 2003).

C2 The ensemble of measurement functions can represent, in a distributed fashion, an object that is not the optimal stimulus for any of them. This is simply a reflection of the principle of coarse coding, which has been widely used in cognitive modeling (Hinton, 1984; Edelman, 1999), and which has a solid basis in neurobiological data; see, e.g., (Rolls and Tovee, 1995; Vogels, 1999).

C3 The measurement functions need not be generic, and need not combine orthogonally (intuitively, they may partially overlap). This design choice, which we justify on the grounds of learnability and ease of use, is related to the idea of using empirically determined, possibly overcomplete basis sets in signal representation (Chen et al., 1999).

C4 Constituents (fragments) of the spatial structure of the object are bound to each other by virtue of being bound to their proper places in the visual field, obviating at least the location variety of the binding problem (Treisman, 1996).

C5 The representation is effectively systematic, subject to certain conditions (Definition 3, appendix A). Specifically, it has the ability, as a matter of principle, to recognize as such objects that are related through a rearrangement of middle-scale parts (as in Hummel's example).

C6 The representation is not compositional in the classical Fregean sense, mainly because fragments of images need not constitute a minimal, orthogonal basis (cf. appendix A). At the same time, our scheme *does* describe composite objects in terms of their spatial constituents (image fragments that are the members of the set of measurement basis functions).

A pilot implementation of this scheme has been described in (Edelman and Intrator, 2000); the example described next is designed specifically to explore aspects of systematicity that are at the core of the present discussion.

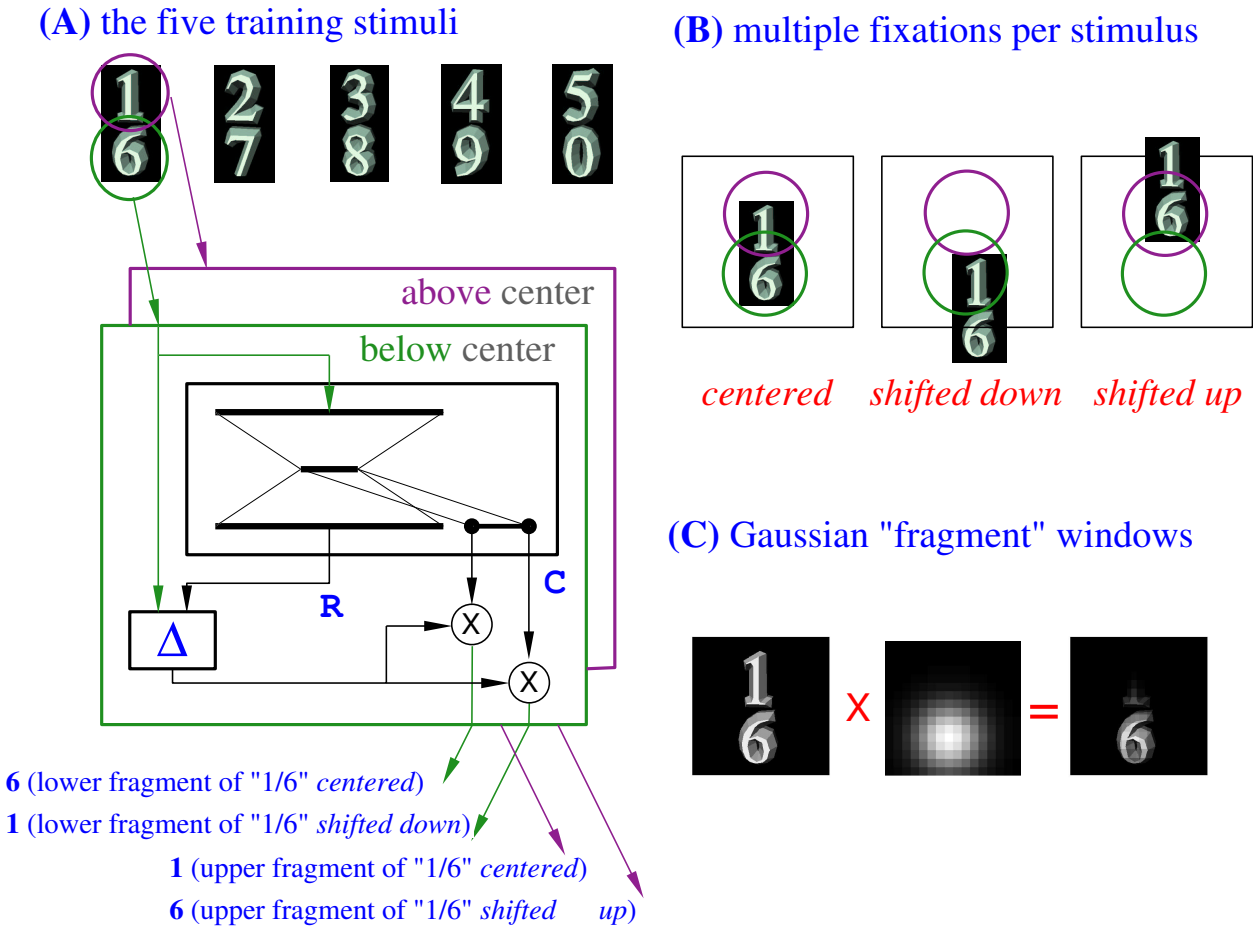

Figure 5: *Left (A):* a 10-unit CoF model, only two of whose units are shown explicitly. The two *what+where* units, labeled `below center` and `above center`, are responsible, respectively, for the bottom and the top fragments of the input image. The model is trained on five composite objects consisting of the 10 digit shapes (1/6, 2/7, 3/8, 4/9, 5/0). Each unit is trained for two-way classification of its input fragment; due to the unary coding of the classification results, this makes for 20 output dimensions for the entire model. Note that the training set is limited in that each of the 10 digit shapes appears always in context either above or below the center of the composite "object." An important measure of the model's systematicity is its ability to transcend this limitation (as illustrated by Figures 8 and 9). *Right top (B):* multiple fixations of the same stimulus during training (but not during testing) are required for a systematic treatment of structure. The "below center" units are trained to discriminate between centered and shifted-down inputs; the "above-center" units are trained to discriminate between centered and shifted-up inputs. Knowledge of the shift of the fixation point relative to the center of the object can be provided during training by an efferent copy of the gaze control information (Arbib, 1979). *Right bottom (C):* fragments are defined by non-crisp, Gaussian windows corresponding to the spatial receptive fields of the bottom and top units. See section 2.3 for an explanation of the model's architecture and training regimen.

2.3 An implementation of the model

The implementation of the CoF model described here (see Figure 5) relies on *what+where* units, which respond selectively both to stimulus shape and to its location (Rao et al., 1997; Op de Beeck and Vogels, 2000). The model uses Gaussian windows to roughly address shape fragments in image-centered coordinates. During learning, it relies on multiple fixations to train multiple versions of the same-shape ("what") unit for different locations ("where"); a single fixation of the stimulus suffices for interpreting it during testing. The model operates directly on gray-level images, pre-processed by a simple simulation of the primary visual cortex (Heeger et al., 1996), with complex-cell responses modified to use the MAX operation suggested in (Riesenhuber and Poggio, 1999).

Each of the 10 *what+where* units is trained, on multiple fixations of a single composite object, to satisfy jointly two criteria: (1) to discriminate among the two relevant positions of the object relative to fixation (the "below center" units are trained to discriminate between centered and shifted-down inputs; the "above-center" units are trained to discriminate between centered and shifted-up inputs; see Figure 5, left); (2) to provide an estimate of the reliability of its output, through an autoassociation mechanism attempting to reconstruct the stimulus image. The learning mechanism (whose outputs are labeled in Figure 5, left, as R and C, for Reconstruction and Classification) is implemented as a radial basis function network (Matlab procedure newgrnn). The two C-outputs of a unit provide a two-dimensional, coarse coded representation of any shape sufficiently similar to either of the two fragments (digits) of its composite training object; cf. (Edelman and Intrator, 1997; Edelman, 1998). The reliability estimate (in Figure 5, left, this is the reconstruction error Δ, which modulates the classification outputs) carries information about category, allowing the C-outputs for objects from radically novel categories to be squelched (Pomerleau, 1993; Stainvas and Intrator, 2000).

In the present implementation of the CoF model, one *what+where* unit is assigned to the top and one to the bottom fragment of each object given to the system. Because each of the 10 *what+where* units is trained for a two-way classification, the resulting representation of a composite stimulus has 20 dimensions, as illustrated in Table 1. The manner in which these units get to process various portions of the input image involves two assumptions.

Multiple fixations. First, we assume that during learning the system performs multiple fixations of the target object, effectively providing the *what+where* units with a basis for interpolating the space of fragment translations (Figure 5, right top). To simulate multiple fixations, the training set contains 100 versions of the images, shifted by $[dx, dy]$, where dx and dy are uniformly distributed in the range ± 6 and $\pm 20 \cup 128 \pm 20 \cup -128 \pm 20$, respectively (image size is 256 pixels).[4] Note that the distribution of dy corresponds to fixations that focus roughly on the center and on the top and the bottom edges of the image.

Gaussian windows. Second, we assume that each *what+where* unit is given as input the entire stimulus image multiplied by a Gaussian gain profile (Figure 5, right bottom).[5] It is important to realize that this mechanism can only address the problem of configurational systematicity when coupled with some means of treating a fragment equivalently when it appears in different regions of the visual field. In the CoF model, this functionality is provided by the multiple fixations performed during the training phase (the implications of this assumption are discussed in section 5), and by interpolation between data acquired in different fixations.

[4] It should be possible to trade off the number of fixations for better tolerance against minor translations of the input.

[5] Window-like gain fields, first described in the context of a single-cell study of cortical area V4 (Connor et al., 1997), have been used in the computational modeling of translation-invariant object recognition (Salinas and Abbott, 1997) and of structure representation (Riesenhuber and Dayan, 1997). Here, however, the role of the window is solely to let a *what+where* unit process, instead of the entire input image, a fragment defined (in image-centered coordinates) by a Gaussian receptive field.

```
>> out=learn_CoF_rbf_ver3({E,'Chars/Nums/1-above-7.iv',1,1,0.6,0.1,1},...
       'run');

 ABOVE, 1-above-6:  AT the CENTER -> 0.120825;  shifted   UP -> 0.000016

 ABOVE, 2-above-7:  AT the CENTER -> 0.000000;  shifted   UP -> 0.000000

 ABOVE, 3-above-8:  AT the CENTER -> 0.000000;  shifted   UP -> 0.000000

 ABOVE, 4-above-9:  AT the CENTER -> 0.000000;  shifted   UP -> 0.000000

 ABOVE, 5-above-0:  AT the CENTER -> 0.000000;  shifted   UP -> 0.000000


 BELOW, 1-above-6:  AT the CENTER -> 0.000000;  shifted DOWN -> 0.000000

 BELOW, 2-above-7:  AT the CENTER -> 0.102116;  shifted DOWN -> 0.008013

 BELOW, 3-above-8:  AT the CENTER -> 0.000001;  shifted DOWN -> 0.000033

 BELOW, 4-above-9:  AT the CENTER -> 0.000000;  shifted DOWN -> 0.000000

 BELOW, 5-above-0:  AT the CENTER -> 0.000000;  shifted DOWN -> 0.000000


 WINNER ABOVE: 1
 WINNER BELOW: 7
```

Table 1: The output of the Matlab function implementing the CoF model, given the image of 1 above 7 as stimulus (cf. Figure 8, left). The 20 dimensions of the distributed representation of this stimulus are similarities to the various localized fragments, such as "bottom of 2-above-7." Note that this happens to be the model's only notion of the shape "7" which is never seen in isolation. The winner ABOVE is "1-above-6 at the center", and the winner BELOW is "2-above-7 at the center," corresponding to the final interpretation "1 above 7."

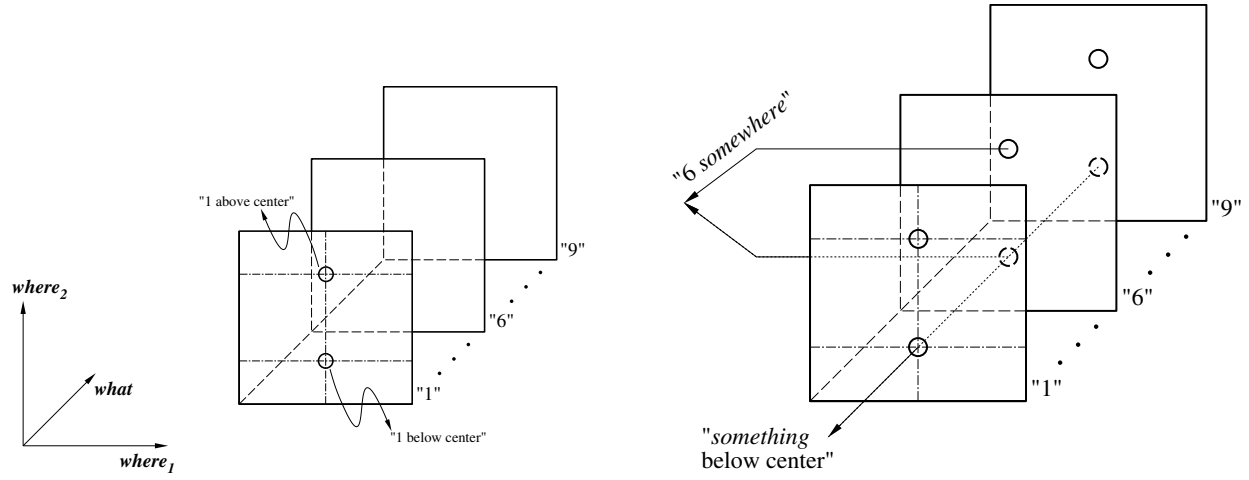

Figure 6: *Left:* the CoF model conceptualized as a "computation cube" with two location dimensions (parallel to the picture plane) and a number of shape ("what") dimensions, represented by processing planes stacked behind each other in spatial register (three such planes are shown, tuned to the shapes of digits 1, 6, 9). *Right:* to determine that, say, a "6" is present in the image, the activities of units along the space dimensions of the computation cube need to be summed, or, better yet, subjected to a maximum detection (Riesenhuber and Poggio, 1999). This signal can then be fed back into the cube to determine the location of the active units, supporting systematic treatment of structure; see Figure 7 and section 2.4. The summation can be also seen as the estimation of the marginal probabilities of shapes and of their locations, which is useful for unsupervised learning of a shape "alphabet" (Barlow, 1990; Edelman and Intrator, 2001).

2.4 Experiment 1: systematicity

The distributed representation of structure by an ensemble of *what+where* units, depicted in Figure 4, is systematic by design, to the extent that the measurement functions implemented by the units can support discrimination among the relevant objects irrespective of their location (see the appendix for a formal derivation of this conclusion). Intuitively, systematicity with respect to Fodor's example – "John loves Mary" – requires that both "John" and "Mary" be identifiable when appearing in each of the two argument slots of the "love" predicate. Likewise, in the visual example illustrated in Figure 1 (right top), the circle and the square shapes have to be identifiable when appearing either at the top or at the bottom of a composite object.

A distributed model of systematicity would be incomplete, however, unless a mechanism is specified that would allow it to realize that the same shape appearing on top of one object is at the bottom of another object. In other words, a mechanism is needed that would be able to compare two distinct subsets of the dimensions of the distributed representation (namely, the outputs of the *what+where* units for which *where*=top with those for which *where*=bottom; see Figure 4).[6] Such a comparison is trivially easy in a system that allows dynamic binding, arbitrary routing, or random-access memory addressing. Our approach to systematicity is, however, motivated by a desire to avoid postulating such operations, which are suitable for an exercise in computer engineering, but are difficult to deal with in the context of a biological model. In the absence of

[6]We thank John Hummel for insisting upon the need for an explicit mechanism of this kind. This point can be related to much wider issues such as the nature of distributed representations, the need, if any, for "binding" in phenomenal experience, and, ultimately, the so-called unity of consciousness. These issues are outside the scope of the present paper; for some relevant observations, see (O'Brien and Opie, 1999; Edelman, 2002).

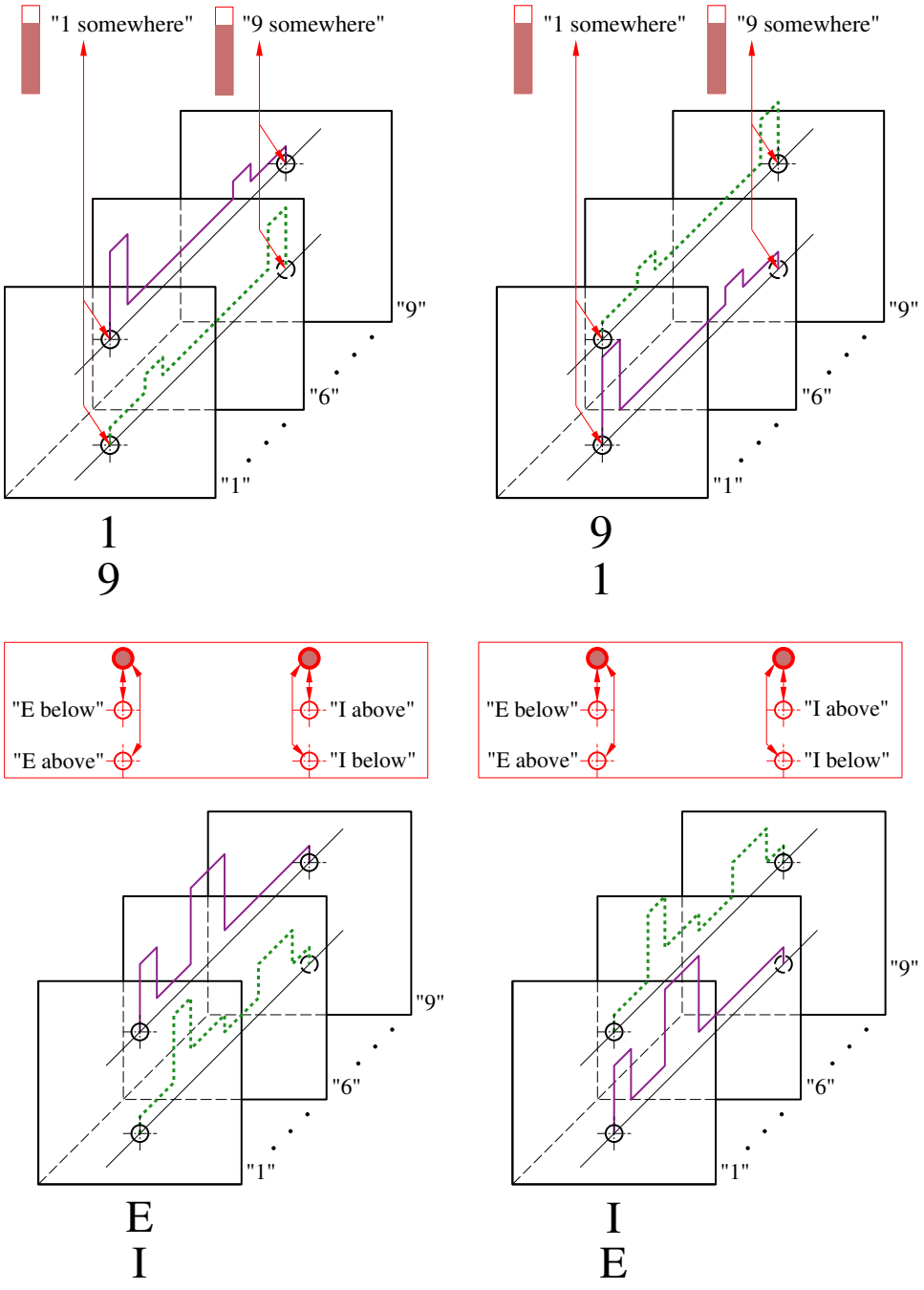

Figure 7: The limited systematicity of the CoF scheme; see section 2.4 for a detailed explanation. *Top:* systematicity for localized representations (objects composed of familiar fragments) is relatively easy. *Bottom:* systematicity for distributed representations is likely to be more difficult, and may require that localized representations of the relevant fragments be created first.

dynamic routing, the (static) model can still determine that two composite objects share fragments through a two-stage procedure:

1. Compute marginals to detect presence of fragments (see Figure 6). The "where" information is lost in this operation; in Figure 7, which offers a schematic example of the representations of 1/9 and 9/1, the ascending signals do not distinguish between these two stimuli.

2. Trace the fragments back to the "what+where" layer, to determine their locations. Although this operation requires descending (feedback) connections, these can be hard-wired; no dynamic binding or routing needs to be postulated. In Figure 7, top, a descending query will activate the "what+where" units responsible for the instances of "1" in 1/9 and 9/1.

The real challenge for this approach arises when viewing radically novel composite objects, for which the representations are likely to be highly distributed (Figure 7, bottom). The issue is the difficulty of computing marginals over distributed representations. The solution we offer is to marginalize (1) over units spanning the dimensions of the coarse-coding space (Figure 7, top), and (2) over any combination of these that merits a dedicated higher-level unit (Figure 7, bottom). This approach predicts that the perception of a structured object that does not merit a dedicated unit (i.e., an object whose representation remains highly distributed) will not be invariant under translation; nor will it be amenable to priming. We remark that some evidence compatible with these predictions is already available (Nazir and O'Regan, 1990; Dill and Edelman, 2001); probabilistic criteria that can help determine whether or not a structured stimulus merits a dedicated representation are discussed, e.g., in (Barlow, 1990; Edelman and Intrator, 2001).

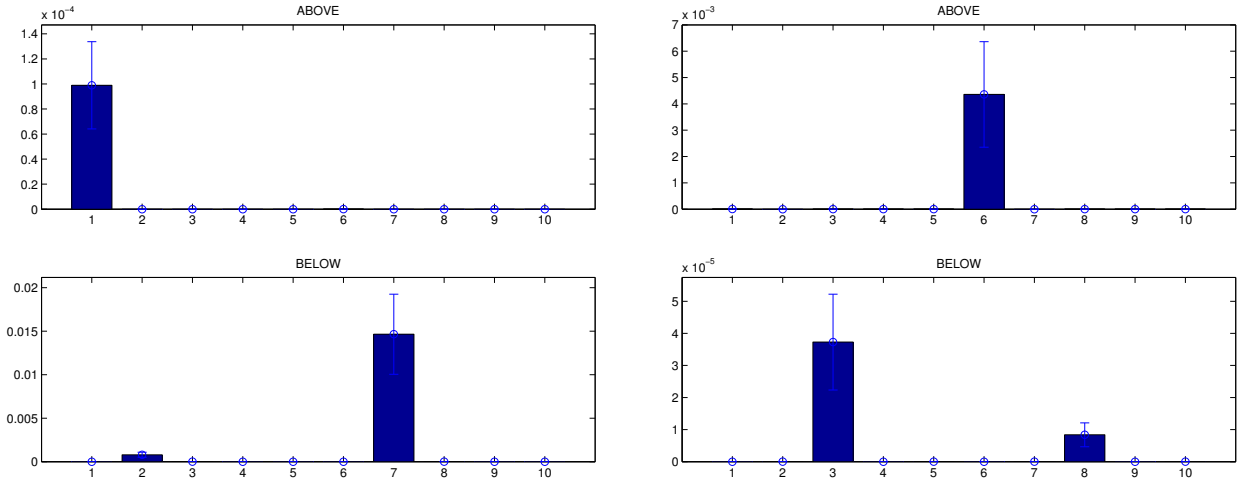

Figure 8: An example of the systematicity of the CoF model on familiar fragments. *Left:* the response of the model to a novel composite object, 1 over 7; the bars show the means $\pm$ standard error of 30 runs. The vertical location of the composite object was distributed normally around the center of the image, with $\sigma_{dy} = 5.0$ pixels. The interpretation was correct in all 30 trials. *Right:* the response of the model to a novel composite object, 6 (which only appeared in the bottom position in the training set) over 3 (which was only seen in the top position). Here too the interpretation was correct – and the model's behavior systematic – in all 30 trials.

Combining the mechanisms illustrated schematically in Figures 6 and 7 with the implemented CoF model, which uses 10 *what+where* units, makes it feasible for the system to address the possible locations

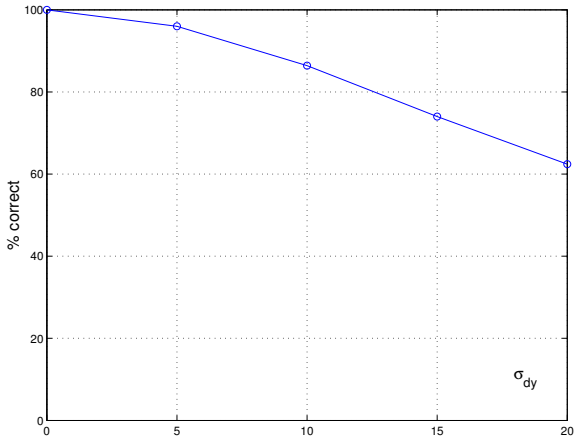

Figure 9: The systematicity of the CoF model on all possible combinations of the familiar fragments (the 10 digit shapes), plotted against the spread constant σ_{dy} of the normal distribution in the vertical location of the composite object. For $\sigma_{dy} = 0$, the interpretations offered by the model were correct in all 100 possible test cases (10 digits on top $\times$ 10 digits on the bottom). Each test run was repeated 10 times.

of a stimulus fragment (digit shape) in the digit pair examples. The resulting system responded largely systematically to the learned fragments even when these were shown in novel locations relative to the centroid of the composite stimulus (Figures 8 and 9). Indeed, given that systematicity may be more difficult to achieve for some sets of stimuli than for others, it makes sense to treat it as a graded rather than an absolute characteristic. One way to define graded systematicity is by considering the proportion of the "total vocabulary" of inputs for which it holds (Hadley, 1994; Niklasson and van Gelder, 1994). This definition can be easily adapted for use with visual stimuli: Figure 9 shows the proportion of the total number of digit pairs which the CoF model interpreted correctly in experiment 1.

2.5 Experiment 2: productivity and systematicity

Our second experiment aimed to demonstrate that the CoF model is productive, in that it can behave systematically on structured objects composed of fragments it never saw during training. We used pairs of uppercase letter shapes as stimuli, submitting two composite objects related by a letter swap to the model and comparing the resulting representations, as illustrated in Figure 10. As before, each representation here is a 20-dimensional vector; because the letter pair objects are unfamiliar to the model, their representations tend to be more distributed than those of digit pairs.

The systematicity of these representations was assessed by computing the correlation between the first 10 dimensions of the representation of one composite object (say, E over I) and the second 10 dimensions of the representation of its counterpart (I over E), and vice versa. Note that the computation of this correlation involves bringing together different parts of the distributed representation vector; this is sanctioned by the assumption of the existence of the hardwired routing mechanism depicted in Figure 7. To simplify the computation, we first swapped the it above and *below* halves of the representation of the first object, then computed the Pearson correlation between the two 20-dimensional vectors corresponding to the two objects. This computation was performed twice for each letter pair ($1352 = 2 \times 26^2$), to allow for the possible asymmetry between the two groups of *what+where* units of the CoF model. The average correlation was 0.71 ± 0.19 (mean and std. dev.); in comparison, the correlation obtained with 1352 random 20-dimensional

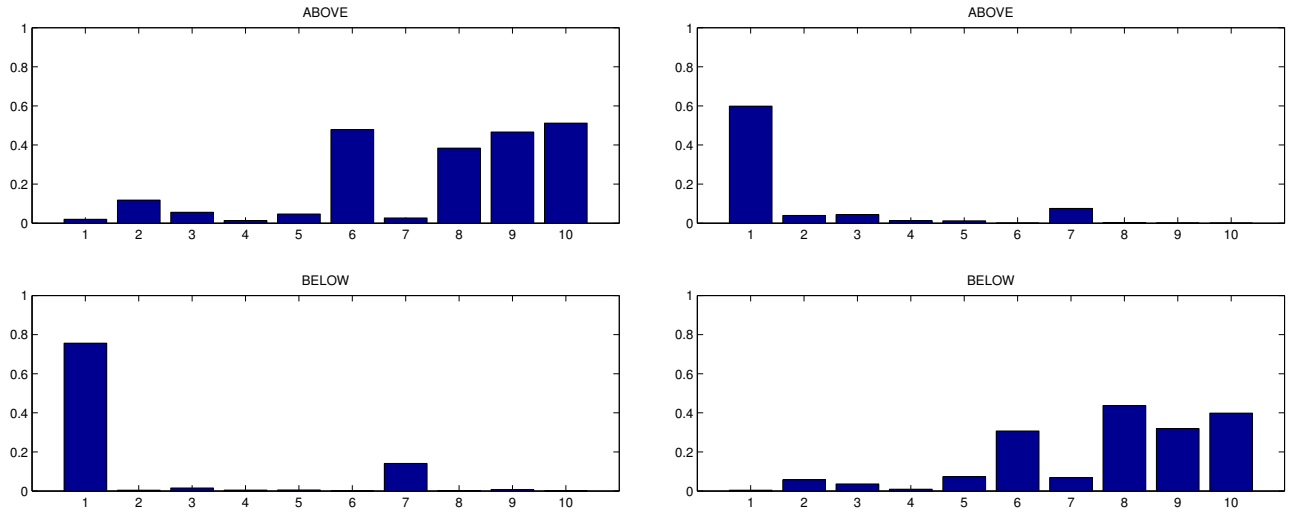

Figure 10: A quantitative assessment of the systematicity of the CoF model on unfamiliar shapes. A model that has been trained on five pairs of digits is shown responding to two pairs of capital letters: E over I (left) and I over E (right). Systematicity is defined as the correlation between the top row of the representation on the left and the bottom row on the right and vice versa, as in Figure 4. The correlation in this trial was 0.97; see section 2.5 for details.

vectors was 0 ± 0.23.

Another way of assessing both the systematicity and the productivity of the CoF model is to examine its representation of the letter pair objects for their ability to support categorization of single letters. A visual impression of this ability can be obtained from Figure 11, which shows the two-dimensional layout of the (20-dimensional) CoF representation space, generated by multidimensional scaling, or MDS (Shepard, 1980). Interestingly, the result resembles the configuration derived by MDS from human letter discrimination data (Gilmore et al., 1979). More importantly for the present study, the representation of letters in terms of similarities to digits is actually good for categorization: the simplest off-the-shelf learning module (Matlab procedure newpnn), trained to a 100% correct performance on half the data, yielded a 85.3% correct performance on the other half of the data. The only parameter required by newpnn, the basis function spread constant σ, was set to 0.01 following a simple cross-validation search.

3 Psychological aspects of visual structure representation

We now proceed to examine the *what+where* approach to systematicity in the light of some of the known characteristics of human behavior in structure-related tasks. Rather than attempting a review of the "object recognition" literature, we concentrate on specific findings that can be brought to bear on the critical assumptions incorporated into the CoF model: the interplay between shape and location information, the role of multiple fixations, hierarchical analysis of shapes, and, generally, the degree of systematicity of the representations formed by human observers.

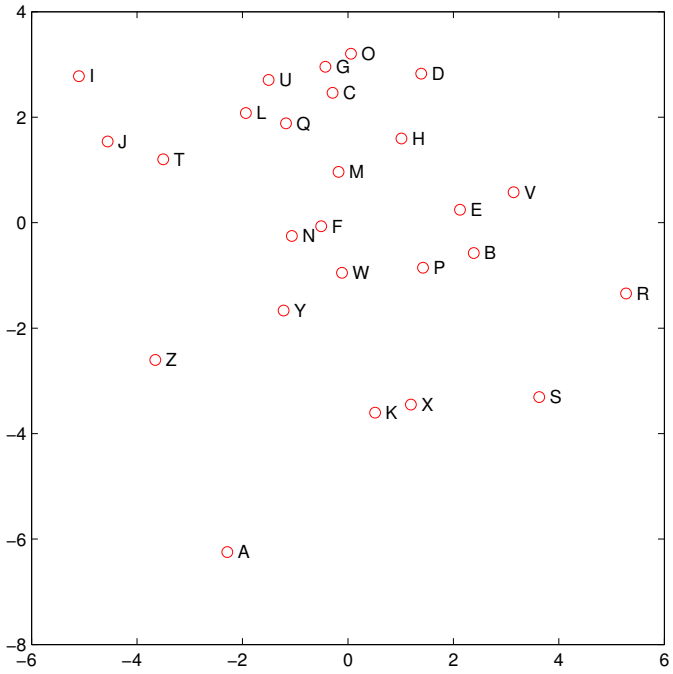

Figure 11: The productivity of the CoF model. In this experiment, the representations of (pairs of) letters by similarities to (pairs of) digits have been examined to determine their ability to support letter categorization. The configuration of the 26 letters in the original 20-dimensional representation space has been embedded into two dimensions for visualization purposes, using metric multidimensional scaling (MDS).

3.1 Limited systematicity

There is no doubt that human observers are capable of perfectly systematic behavior towards composite objects such as "circle above square" (see Figure 1, right top), which can be considered a visual counterpart of the "John loves Mary" example typically encountered in discussions of systematicity. In such stimuli, every available cue, from Gestalt principles, through perceptual experience, to abstract knowledge, points toward a decomposition into simple geometrical parts. It is known, however, that human observers are not systematic in their perception of upside down images of objects such as faces, which possess a highly overlearned "normal" orientation (Rock, 1973; Thompson, 1980). We suspect that humans are also less than systematic with complex scenes such as those in Figure 1, left, unless given an opportunity to scrutinize them, while modulating both the spatial scale and the location of the focus of attention, and bringing the full power of conceptualization and symbolic reasoning to bear on the visual stimulus.[7] Although visual systematicity for complex stimuli has not been studied to date, many available results indicate that observers are limited in certain ways in their ability to represent even relatively simple spatial structures.

The role of multiple fixations. The first such limitation, which holds also for the CoF model, is the apparent need for multiple fixations of the stimulus, discussed at length by (O'Regan and Noë, 2001). To cite an example, (Schlingensiepen et al., 1986) found that without eye movements, observers had difficulty dis-

[7] In sentence comprehension, subjects are indeed less than systematic with stimuli such as "The girl who Paul saw after discovering Alex proposing to dismiss had lunch in a cafe" (Chipere, 1997).

tinguishing patterns composed of arrays of random black and white squares. Likewise, (Nazir and O'Regan, 1990) trained subjects to distinguish among several shapes resembling Chinese characters under conditions that prevented eye movements. They found that hundreds of trials were required even to discriminate between the stimuli, and that observers often failed to recognize the learned patterns at a new retinal location, only half a degree away from the familiar position.

Imperfect translation invariance. More recently, (Dill and Edelman, 2001) found that performance in a same/different discrimination task using articulated animal-like 3D shapes was fully transferred across retinal location if local cues were diagnostic, but not if the decision had to be based on relative location of various fragments. The findings of that study can be summarized as follows: translation and size invariance hold for 3D stimuli (both familiar and unfamiliar) that can be discriminated on the basis of local shape information (e.g., the thickness of this or that limb of the animal-like shape), but not for stimuli whose only distinguishing cues are configurational or structural. This pattern of results suggests that this mechanism treats local cues and structural information differently, in a manner that is compatible with the predictions of the CoF model. Indeed, in an ensemble of *what+where* units the relational structure of the stimulus is represented implicitly, in a distributed fashion. Unlike individual local features, structure would not, therefore, be amenable to translation-invariant priming — as indeed reported by (Dill and Edelman, 2001).

The role of attention. In some situations, the representation of spatial relations requires spatial attention. For example, the experiments of (Logan, 1994) examined the role of spatial attention in apprehending the relations *above*, *below*, *left*, and *right*. Logan found that visual search was difficult when targets differed from distractors only in the spatial relation between their elements, suggesting that attention is needed to process spatial relations. Support for the idea that the grasping of spatial structure[8] is not automatic, but rather needs the mediation of attentional mechanisms comes from the work of (Wolfe and Bennett, 1997), who referred to preattentive representations of objects as "shapeless bundles of basic features." Likewise, a review of the literature conducted by (Treisman and Kanwisher, 1998) suggests that "attention is required to bind features, to represent three-dimensional structure, and to mediate awareness."

The role of context. The classical, symbolic/propositional approach to the representation of composite visual objects (Hummel and Biederman, 1992; Hummel, 2000) predicts context-independent performance in structure-related tasks. In contrast, the perceptual symbol systems theory (Barsalou, 1999), which posits image schema-like representations (not unlike the "corkboard" behind the CoF model), predicts context sensitivity: the same visual feature would be "local," hence potentially variable across concepts. Indeed, this is what has been reported: "different local representation of *wings* exist for ROBIN, BUTTERFLY and JET, each capturing the specific form that *wings* takes in its respective concept" (Solomon and Barsalou, 2001).

3.2 Interdependence of shape and location cues

An intimate connection between the representation of visual space (location in the two-dimensional visual field, or the "where" information) on the one hand, and the representation of shape (the "what" information) on the other hand is a basic tenet of the CoF model. At least in some recognition-related tasks, "what" and "where" cues are indeed intertwined, as indicated by the findings of (Wallach and Austin-Adams, 1954).

[8] In the terminology of our model, this would correspond to "marginalization"; see Figure 7 and the accompanying explanation.

These researchers found that the interpretation of an ambiguous shape could be biased by priming with an unambiguous version, but only if both appeared within the same visual quadrant. A similar confinement of the priming effect to a quadrant was found, in a subliminal perception task, by (Bar and Biederman, 1998); for other evidence of structuring of spatial categorization by quadrants, see (Huttenlocher et al., 1991; Crawford et al., 2000). Interestingly, a lesion in V4, the highest area in the ventral stream where quadrant information is still relatively well separated, can result in the subject being able to perceive the constituents of a composite object, yet finding it difficult to determine their configuration (Gallant et al., 2000).

The notion that the representation of an object may be tied to a particular location in the visual field where it is first observed is compatible with the concept of *object file* — a hypothetical record which is created by the visual system for every encountered object and which persists as long as the object is observed (Kahneman et al., 1992). Results obtained by Treisman and her associates, summarized in (Treisman, 1992), indicate that "location" (as it appears, e.g., in the CoF model) should perhaps be interpreted relative to the focus of attention, rather than retinotopically (more on the role of attention below).

(Cave et al., 1994) report an investigation of the way location is represented in the visual system. The subjects in their study performed mental rotation of stimuli whose location varied from trial to trial. In one of the experiments, distance between stimulus locations was varied systematically. Response time increased with distance, suggesting that image representations are location-specific, In another experiment, the subject had to make an eye movement to fixate the test stimulus. The subjects responded more quickly when the test stimulus appeared at the same retinotopic location, not the same spatiotopic (allocentric) location as the cue, suggesting that location is coded retinotopically in image representations.

Conjunctions of shape and location play a central role in the CoF model: such conjunctions are precisely the kind of stimulus the "what+where" units are supposed to be tuned to. (Saiki and Hummel, 1996) examined the representational status of conjunctions of part shapes and relative locations, showing that in object category learning subjects are particularly sensitive to these complex features. Participants in their study learned categories defined by a part's shape and color (part-color conjunctions) or by a part's shape and its location relative to another part (part-location conjunctions). The subjects were found to be better at classifying objects defined by part-location conjunctions than objects defined by part-color conjunctions.

(Johnston and Pashler, 1990) studied the binding of identity and location information in disjunctive rather than conjunctive feature search. Subjects searched a stimulus display for a color or a shape target, and reported both target identity and location. The results of these experiments indicated a strong binding of identity and location; in fact, no perception of identity without location was found (location, in contrast, did seem to be represented to some extent independently of identity). A similarly central role of location in defining the stimulus has been reported by (Shapiro and Loughlin, 1993), who used the negative priming effect to investigate the nature of "object files" containing both identity and location information.

It is interesting to observe that subjects seem to perceive relative location as a graded rather than categorical cue (Hayward and Tarr, 1995). Following a linguistic analysis of the use of spatial prepositions, (Talmy, 1983) suggested that their meanings collectively cover the range of possibilities (i.e., of the spatial relations that need to be encoded) much the same way graded overlapping receptive fields cover the visual space. Hayward and Tarr examined this idea experimentally, by mapping the spatial extent of regions in the visual field in which one had to place objects so as to satisfy the subjects' idea of the meaning of *above*, *below*, etc. They found that the accuracy of position estimates and the sensitivity to shifts in position varied, and were both highest when the target object was in a spatial location where spatial terms had been judged to have high applicability. These results indicate that the structure of space as encoded by language may be determined by the structure of spatial relations in visual representation; they also support the idea of graded,

non-categorical representation of relative location, inherent in models such as CoF.

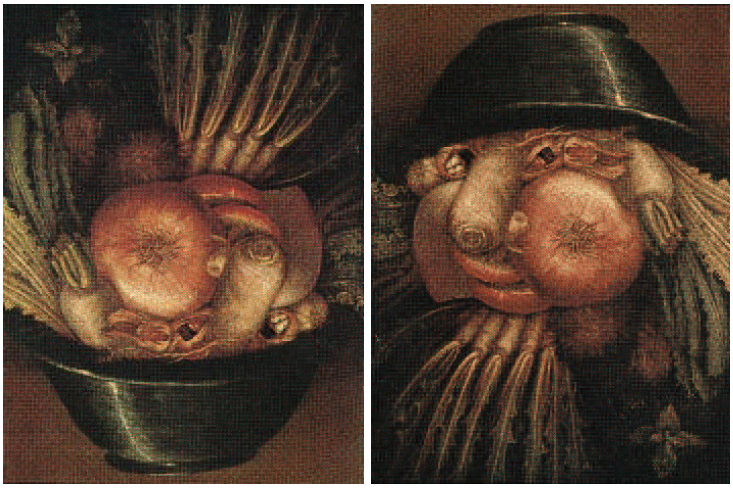

Figure 12: *The Cook*, a painting by Giuseppe Arcimboldo. On the right, attention seems to be drawn first to the global level of description (a face). Only subsequently, a breakdown of the gestalt and the concomitant perception of constituent patterns, if any, may occur. Moreover, it seems to be difficult to attend simultaneously to more than one spatial scale: either the vegetables, or the face, but not both, dominate perception at any given time. On the left, there is no compelling gestalt, and no rivalry between coarse and fine scales. Section 3.3 lists experimental data that support these intuitive observations.

3.3 Global precedence and shallow structure

Much of the current interest in the perception of structures that span multiple spatial scales has been inspired by the work of (Navon, 1977), who argued for the precedence of global information (hence, large-scale structure) over local (cf. Figures 12 and 13). Since then, the idea of precedence of global cues has withstood extensive testing. For example, (Sanocki, 1993) describes integration priming experiments, in which primes and targets were presented briefly, then masked. It was found that global, coarse-grained common-feature primes presented early in processing facilitated discrimination between similarly shaped targets, even though they provided no discrimination-relevant information. Moreover, global primes were more effective than local ones early in processing, and this situation was reversed late in processing. The importance of attending to the right level of structure is underscored by the work of (Venturino and Gagnon, 1992), who found, using slides of natural scenes, that subjects are slower and less accurate when their attention and the forced-choice alternatives dictated by the discrimination task are at different levels of stimulus structure.

In another investigation of this issue, (Love et al., 1999) tested the idea that "structural relations among elements" influence the relative speeds of global and local processing in a same/different decision task. With the perceptual salience of the global and local structure equated, advantages were still found in the processing of global shape. In particular, Love's subjects were able to process the relations among the elements quickly, even before the elements themselves were identified.

4 Neurophysiological aspects of visual structure representation

Neurobiological evidence is particularly germane in deciding the plausibility of the coarse coded, statically bound approach to the representation of structure, given the origins of the concept of *what+where* cells in a study of the primate prefrontal cortex (Rao et al., 1997). In this section, we survey neurobiological data that roughly parallel the behavioral findings mentioned above.

Neuronal mechanisms that can support the shape selectivity function of the units in the CoF model have been found in the inferotemporal cortex by (Logothetis et al., 1995). Most of these cells respond selectively to particular views of an object; responses of such cells can be combined to support selectivity to a specific shape, largely irrespective of view. Numerous other reports of face and object selectivity are reviewed, e.g., in (Logothetis and Sheinberg, 1996; Rolls, 1996; Tanaka, 1996; Edelman, 1999).

Evidence concerning selectivity of the shape-tuned cells in the higher visual areas such as TEO and TE to stimulus location has only recently begun to surface. For decades, the field has been dominated by the idea of separate ventral ("what") and dorsal ("where") streams, derived initially from primate lesion studies (Mishkin et al., 1983). According to this notion, cells at successive stages of the "what" stream are supposed to have larger and larger receptive fields, culminating in absolute invariance with respect to translation.

The concept of translation invariance is closely related to that of configurational systematicity. In the CoF model, two objects related through translation will be represented by different measurement functions; if the measurements conform to Definition 3 (given in the appendix), the representation will be systematic, yet the identity of the two objects would only become apparent through the expenditure of an additional computational effort (e.g., through associative learning and subsequent generalization; cf. Figure 7 and (Hadley, 1997), p.145.). Systematicity, therefore, can be seen to stop short of requiring translation invariance, as it only should: if the structure of an object is to be made explicit, it would not do to lose track of where exactly each of its fragments is located. Not surprisingly, schemes that start by extracting translation-invariant representations of objects — including a two-stream visual system that separates "what" from "where" — are susceptible to the binding problem (von der Malsburg, 1995; Treisman, 1996), of the kind that our model addresses through the use of the visual field itself as the scaffolding (or corkboard) for structure.

The separation between the streams has been questioned in a more recent survey (Ungerleider and Haxby, 1994), in view of the many examples of information interchange. Moreover, cells of the *what+where* variety, which are tuned *both* to shape and to its location in the visual field, have also been found, in areas V4 and TEO by (Kobatake and Tanaka, 1994), and in the prefrontal cortex by (Rainer et al., 1998); the latter study is the one in which the concept of a *what and where* cell has been coined. A study of the spatial receptive fields of shape-selective cells in the inferotemporal cortex has been undertaken recently (Op de Beeck and Vogels, 2000). The detailed maps yielded by this study indicate that the receptive fields in IT vary in size ($2.8^\circ - 26^\circ$), differ in their optimal position, and are graded (well-approximated by 2D Gaussians). The authors conclude that these cells can code for the location of stimuli in the center of the visual field; we suggest that they can also serve as building blocks in location-bound representations of structure proposed here.

The idea that location binds together, statically, object fragments that belong together is supported also by the lesion study mentioned earlier (Gallant et al., 2000), which highlighted the role of area V4 in the perception of object structure. Another patient study, in which the locus of the lesion was in the pulvinar nucleus, indicates that subcortical areas too may be involved in location-based binding (Ward et al., 2002). In the monkey, the rostral part of the pulvinar is known to contain spatiotopic maps of the inferior and lateral parts of the visual field; indeed, Ward et al. found that their patient had a deficit there in "the use of spatial information in binding."

5 Discussion

In this section, we mention several related computational schemes of structure representation, and discuss a number of open issues, such as learning the fragments, and the need to deal with nested structures.

5.1 Related computational schemes

As we pointed out in the introduction, the Recognition By Components theory of visual structure representation (Biederman, 1987), incorporated into the JIM model of (Hummel and Biederman, 1992), is both productive and systematic. The alternative approach described here was motivated by our desire to simplify the computational (as well as philosophical) assumptions behind the representation of structure, and to base it on uncontroversial characteristics of the primate visual system, such as retinotopy (Edelman, 1994; Edelman, 1999; Edelman and Intrator, 2000). Static binding by retinotopy has been adopted also by the most recent version of the JIM model (Hummel, 2001). It is used there alongside dynamic binding, to process stimuli faster and without need for attention, at the expense of some invariance to object transformations. The base representation for both static and dynamic streams of this model is the same as in the earlier versions (a collection of categorical shape features that feeds detectors for generic part shapes, or geons). The model accepts as input symbolic descriptions of line drawings of 3D objects, thus by-stepping a major hurdle common to the part-based approaches: the need for reliable detection of the parts in raw images (Dickinson et al., 1997; Edelman, 1997).

This difficulty is well known to workers in computer vision, where attempts to represent objects in terms of the arrangement of their constituents typically use image snippets instead of categorical parts, and robust statistical evidence accumulation instead of logical operations for inferring object presence (Burl et al., 1998; Sali and Ullman, 1999; Heisele et al., 2002). We note that the issue of binding does not really arise in a computer vision setting, for the reason stated in the introduction: once the full power of general-purpose computation is assumed, binding becomes trivial. Nevertheless, it is interesting to observe that all the schemes of the kind just mentioned describe structure in a "retinotopic" frame, by coding the image locations of the constituents of an object.

Because the focus of the computer vision methods is on recognition and categorization, the issue of configurational systematicity of the "circle above square" ("John loves Mary") variety tends not to be discussed there. In categorization, context systematicity (recognizing wheels as such in different cars) is a functional necessity, while configurational systematicity is a bonus; it could be useful with scenes, but not with objects, because scrambled objects do not look like anything in particular and cannot be categorized. In fact, even in the field of computational neuroscience of vision, very few models address the issue of systematicity explicitly. For example, a hierarchical competitive model of binding (Elliffe et al., 2002), based on principles similar to those of the Neocognitron (Fukushima, 1988), does not mention systematicity (the stimuli processed by this model are composed of one to four line segments in tightly constrained mutual positions and orientations).

A neural network model capable of learning non-classical, coarse-coded representations that aim explicitly at configurational systematicity has been described by (O'Reilly and Busby, 2002). This model represents objects by distributed patterns of activation over eight features per location within a 4×4 array of locations. Four possible relations are coded (right, left, above and below), and the inputs consist of two objects. The model is trained to answer questions such as "what," "where," and "what relationship"; to do so, it must bind object, location and relationship information. O'Reilly et al. demonstrate the generalization of this capability to novel inputs: a model trained on 10% of the space of possible inputs generalized at

80% correct; training on 25% of the inputs led to a 95% correct generalization. It is difficult to compare those rather impressive figures with our results, because of the differences in input types and generalization sets (our model used gray-level images of bipartite 3D objects, allowing only one kind of relation, vertical stacking; generalization was assessed by training on digit pairs and testing on letter pairs).

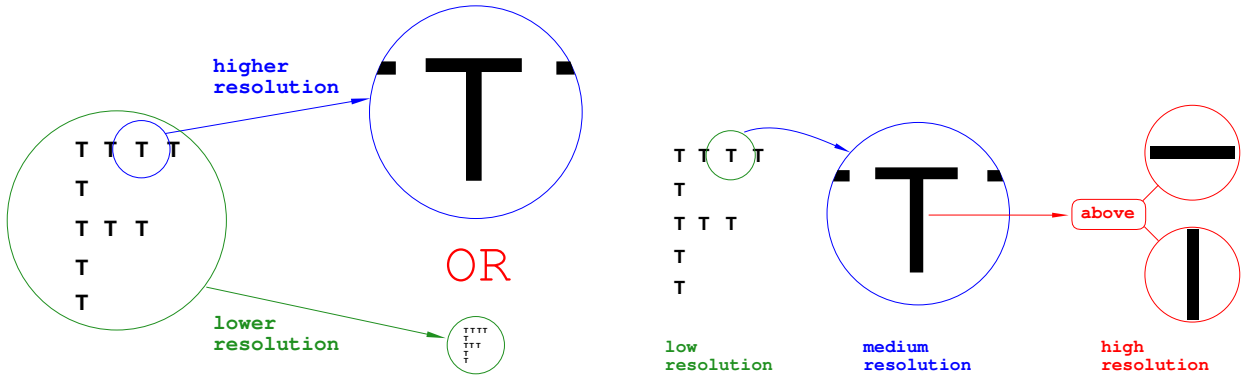

Figure 13: *Left:* shallow-scope compositionality. This illustration shows a large letter F composed of smaller instances of a T — a stimulus frequently used by psychologists in studies of local vs. global processing of visual shapes (Navon, 1977). According to the principle of shallow scope, this stimulus would be described as a large roughly centered F, or as a (small) T in a particular, possibly off-center location, but not simultaneously as an F and a collection of Ts. *Right:* recursively applicable compositionality. To make explicit the structure of the stimulus at a finer scale than currently employed, a system that operates on the principles we postulate must focus attention on the area of interest, and adjust its spatial resolution accordingly. Two successive operations of this kind are illustrated here. We doubt that more than three levels (spatial scales) of attention need to be involved in the analysis of commonly encountered structured objects.

5.2 Hierarchical systematicity

A system such as ours, in which only one structural level is made explicit at any given time, cannot attend simultaneously to a hierarchy of levels. Intuitively, the forest, the trees, and the branches cannot be taken in all together at the same time (see Figure 13). The level of representation maintained by such a system on an instantaneous basis can be controlled via the spatial extent (scale) of the attention window. On the local level, the focus of attention can be directed to the area of interest. If the relevant processing mechanisms are self-similar across scales,[9] this combination of steerable attention and control over spatial resolution will achieve precisely the desired effect: structure on several scales will be treated on an equal footing diachronically (with only a single-scale representation being active at any given time).

The shallow-scope, single-scale-at-a-time approach to representation need not limit the ability of the system to deal with nested structures. Such structures can be described recursively, by means of focusing attention and increasing or decreasing the effective resolution. For example, at the entry level of categorization (Jolicoeur et al., 1984), the human visual system may represent a human shape as comprising a head, a body, arms and legs; the head may be further described as having a mouth, eyes and ears. At a close range, smaller-scale details (e.g., the whites of the eyes) may be discerned; structurally, these would be linked to

[9] As they are in representations based on a wavelet expansion of the data (Strang, 1989; Field, 1993).

other entities at the same scale (e.g., the pupils or the eyebrows), but not across scales (e.g., the neck or the hands).[10]

Curiously, the "classical" systematicity (Definition 1) involves a proposition (the relation aRb) that is structurally "flat" in that it contains no explicitly appearing recursively embedded propositions. Despite its basic reliance on a flat construction, this definition obviously does allow for the representation of multiply nested structure, because each of the variables a, b can stand for an entire formula, e.g., $a = cQd$, implying $(cQd)Rb$. Leaving the recursion implicit is, in fact, a useful idea: if the nested structure were to be made explicit, the definition of systematicity would have to contain a potentially very large number of formulas — all the nodes of the lattice formed by recursive substitution of variables, up to a given depth. We believe that the flat, implicitly recursive definition of systematicity is a good starting point for a future development of a computationally effective scheme for the representation of hierarchical structure.

5.3 Learning the measurement functions

In this paper, we assumed that the scale of the fragments in terms of which images were to be represented had been given to the system in advance, and, moreover, that all the fragments at that scale were to be encoded (recall that the fragments seen during training were pairs of digit shapes, each of which was half the size of the entire object). In reality, it would be up to the model to figure out that objects may be composed of recurring fragments, and to self-organize in a manner that would allow it to deal with novel configurations of those fragments (or with radically new inputs). This problem in unsupervised learning is, however, outside the scope of the present paper. We show elsewhere how it can be addressed within a statistical inference framework, and provide psychophysical and computational evidence that helps elucidate the role of various criteria of statistical independence in this learning task (Edelman et al., 2002; Edelman et al., 2003). A future work along these lines should probably also consider methods for learning distributed representations of object transformations (Zemel and Hinton, 1995), and for evidence combination and distributed control, both linear (Jacobs et al., 1991) and nonlinear (Lewicki and Sejnowski, 1998).

5.4 Predictions for psychophysics and neurobiology

The functional and the architectural assumptions incorporated into the CoF model generate specific predictions that can be examined empirically by behavioral, electrophysiological and neuroanatomical methods. For psychophysics, the first prediction is that multiple fixations over a radically novel structured stimulus are needed if a rearrangement of its constituents is to be treated systematically. Second, assuming that the "windows" through which *what+where* units see the world are Gaussian, we predict that the systematicity will be more limited for more highly interpenetrating non-convex fragments, even with controlled Gestalt goodness. Third, we predict that a masking study would reveal a timing difference between the early (and largely automatic) awareness of the shapes contained in a composite stimulus, and the late (and more goal-driven) awareness of their locations within the whole – assuming that awareness is a top-down process that reaches the topmost level of structural representation first (cf. Figure 7).

A related prediction holds for single-cell responses in areas V4 and TEO, which can be obtained by electrophysiological means: same-area cells with larger spatial receptive fields should have longer latencies relative to stimulus onset – assuming that such *what* cells are upstream from *what+where* cells with smaller

[10] In a discussion of the hierarchy of terms for various body parts in English, Langacker notes the "nonexistence and oddity" of expressions like **bodytip* or **facelash* (compared to *fingertip* and *eyelash*), and the infelicity of sentences such as *?An arm has 14 knuckles and 5 nails* (compared to *A finger has 3 knuckles and 1 nail*); see (Langacker, 1990), p.8.

RFs. For the latter, the peri-stimulus time histograms should be bimodal, with the first mode corresponding to the response of the cell to the bottom-up signal, and the second to the combined effect of bottom-up signal and top-down priming. Our second prediction for electrophysiology states that the multidimensional ensemble response of *what+where* cells should contain rich information about the spatial structure of the stimulus; when separated, say, into two populations, corresponding to the upper and the lower hemifields, the response patterns should be systematic for stimuli related by spatial transformations such as those of Figure 1.

Finally, we are intrigued by the possibility that the dimensions of selectivity postulated by the corkboard theory of binding and by the CoF model may be mapped explicitly onto the functional architecture of the inferotemporal cortex. Cast in neurobiological terms, the scheme of Figure 7 corresponds to an assertion that *what+where* cells with like *what* selectivity properties should be reciprocally wired to cells with the same *what* selectivity and a wider *where* tuning at a higher stage of processing. Although this prediction would be difficult to test at the level of individual cells, it may hold also at the level of a cortical microcolumn, in which case it could be testable by a combination of electrophysiological and anatomical methods (Lund et al., 1993).[11]

5.5 Conclusions

The traditional route to systematicity and productivity — two issues that are indispensable for the understanding of advanced cognition — is via classical, propositional, part-based compositionality (Bienenstock and Geman, 1995; Fodor, 1998). This approach, which is adopted in vision by Biederman's Recognition By Components theory mentioned earlier (Biederman, 1987), and by many others, is problematic, because of the questionable ontological status of the parts it postulates, and because of the binding problem it creates. In this paper, we showed that a representational system can be considerably systematic, without postulating categorical parts, and without resorting to dynamic binding.

An early indication of the emerging central role of systematicity in the debate on the nature of cognitive representations can be found in (Touretzky, 1989). This and many other works questioning the classical notion of systematicity (van Gelder, 1990; Smolensky, 1990; Niklasson and Boden, 1997) focused on the possibility of a "connectionist" alternative: a mode of distributed representation that would not be bound by compositional rules. As a result, the debate is now waged mainly between the proponents of the classical compositional view (such as Fodor) and those who favor some method of connectionist representation that is altogether non-compositional (e.g., by virtue of being context-dependent), yet, in some sense, systematic.

Although we find much of the critique of the classical view offered by the connectionists pertinent and useful, we do not believe that having to choose between the classical Fregean compositionality and a radically non-compositional representation is a good idea. On the one hand, the classical compositional (and therefore systematic) framework is trivially easy to implement on a symbolic computer, but are not so easy for biological neural networks; on the other hand, non-compositional representations, which can be easily learned by artificial neural networks, have difficulties with systematicity. The present paper espouses a middle road between the two extremes corresponding to the two sides in this debate. In the compromise

[11] The columnar structure of the inferotemporal (IT) areas, mirroring in some respects that of the primary visual cortex, emerges both from single-cell studies (Tanaka et al., 1991; Fujita et al., 1992) and from optical imaging of the cortex (Tsunoda et al., 2001). No conclusive data are available to date concerning the make-up of the individual IT column, although there are indications that cells selective to an orderly progression of different views of an object may be arranged in spatial proximity to each other (Wang et al., 1998). It is also not known whether the columns in IT form a larger-scale structure such as the V1 hypercolumn. It would be interesting to find out whether neighboring IT columns contain cells of similar shape selectivity (Wang et al., 2000), but with varying spatial selectivity.

we offer, the compositional framework is modified to gain much needed biological plausibility, by adopting a distributed approach to the representation of the primitives, and a static, spatial basis for their structural binding.

According to our configurational notion of systematicity, a representation is systematic if it can deal equally well with various spatial arrangements of the same "parts." Such behavior has been exhibited by our model, lending support to the claim that distributed representation of primitives, coupled with the corkboard approach to binding is a promising way of dealing with structure. We feel, however, that configurational systematicity is not the only possible formalization of the intuitive concept of a good representation of structure. In the introduction, we mentioned one alternative: context systematicity, which calls for a principled treatment of homologous substructures (heads of animals, wheels of cars) that recur in various larger-scale structures. Although context systematicity seems to be subsumed under the configurational rubric, the former has not yet been defined, and the relationship between these two concepts is unclear. We hope that this discussion will lead to the emergence of a more comprehensive rigorous notion of systematicity, and, eventually, to the development of theories of structure representation that are intuitively acceptable, formally adequate, computationally viable, and, crucially, amenable to implementation in the brain.

Acknowledgments

We are indebted to Rich Zemel for pointing out a problem in an earlier version of the model. Thanks to Shalom Lappin for help with computational semantics, Zoltan Szabo for advice on compositionality, and to John Hummel, Rich Zemel and an anonymous reviewer for constructive comments on earlier versions of this article.

A Compositionality and systematicity

Recent work in computational semantics (a field motivated equally by theoretical linguistics and by practical needs arising from natural language processing) resulted in developments that are directly relevant to the central concern of the present paper: the representation of structure in vision. The ideas surveyed briefly in this appendix suggest that it is possible to represent structure systematically without necessarily adopting the classical compositional approach.

A.1 A formal treatment of compositionality in computational semantics

In its most abstract form, the issue of compositionality is at the focus of attention in computational semantics — the field which can be said to have originated with Frege's work. Recall that according to Frege (1891), a structure is considered compositional if its meaning (interpretation) is a *function* of the meanings of its parts. The following definition formalizes this idea, and leads to some interesting implications.

Theorem 2 (Zadrozny, 1994) *Let M be an arbitrary set. Let A be an arbitrary alphabet. Let "." be a binary operation, and let S be the set closure of A under ".". Let $m : S \rightarrow M$ be an arbitrary function. Then there is a set of functions M^* and a unique map $\mu : S \rightarrow M^*$ such that for all $s, t \in S$*

$$\mu(s.t) = \mu(s)(\mu(t)), \text{ and } \mu(s)(s) = m(s).$$

Essentially, this means that as long as m is a function (that is, a mapping that associates a single value with its argument), the interpretation induced by it will be compositional. According to Zadrozny, "one of the more bizarre consequences of [this Theorem] is that we do not have to start building compositional semantics for natural language beginning with assigning meanings to words. We can do equally well by assigning meaning to *phonemes* or even *letters* [...]."[12] When applied to the problem of object structure representation, this realization entails that even the smallest bits of images — pixels — can serve as a basis for erecting a perfectly compositional edifice. This atomistic approach would be compositional, at the expense of forcing one to assign an interpretation (meaning) to each and every pixel, a prospect which we do not find at all appealing.

As noted by (Zadrozny, 1994), Theorem 2 shows that the compositionality principle is formally vacuous, unless some constraints are imposed on the interpretation function (in the case of computational semantics, on the homomorphism between the structure of an expression and its meaning). In a later work, Zadrozny re-analyzes the concept of compositionality under the following assumptions: (i) that the meaning of a construction be derived from the meanings of its parts *in a systematic way*; (ii) that the meanings of the parts have some intuitive simplicity associated with them, and (iii) that "one way of building compositional semantics be better than another" (Zadrozny, 1999). Interestingly, this approach, which is based on the Minimum Description Length (MDL) principle,[13] parallels recent attempts to develop a compositional framework for image analysis (Geman, 1996; Bienenstock et al., 1997). Note, however, that assumption (ii) implies interpretational atomism and ontological commitment to the reality of meaningful "parts" — two design choices we consider advisable to avoid, for reasons some of which have been detailed in the body of the paper.

In traditional semantics, the assumption that isolated words have well-defined meanings which are then recursively combined (Katz and Fodor, 1963) was found to be problematic (Lakoff, 1971). For example, (Eco, 1976), p.98, points out, *inter alia*, that words have multiple meanings, which, moreover, depend on context. In vision, additional problems can be discerned. For instance, as we already noted, there are shapes that do not admit a natural compositional description: what would be *the* structural decomposition of a loaf of bread, or a shoe (Ullman, 1989)? For such objects, one would have to start with very simple atomic primitives (pixels or edge elements), exacerbating the problem of finding a stable optimal (in the MDL sense) description (Edelman, 1997).

A.2 Relational systematicity

In view of the central role of classical compositionality in semantics, it is especially interesting to note that non-compositional approaches to the computation of meaning are now being considered in that field (Lappin and Zadrozny, 1999):

> [...] it is possible to construct a theory of meaning which is both non-compositional and systematic. This involves taking the meaning of a syntactically complex expression E to be the set of values of a *relation* [our italics] on the meanings of E's syntactic constituents rather than the value of a function.

Lappin and Zadrozny proceed to give an example of such a scheme, in which meaning is effectively construed as a *set of possible interpretations* rather than a single disambiguated interpretation. This interpre-

[12] Zadrozny then remarks: "But then the cabalists had always known it."

[13] The basic idea is to consider the simplest maximal description of the data that satisfies the postulate that the meaning of the whole is a function of the meaning of its parts (Zadrozny, 1999).

tation scheme maintains systematicity by fully preserving the information needed to establish (e.g., via constraint propagation) the links between meanings of related expressions (Nerbonne, 1995; Lappin and Zadrozny, 1999).

A.3 Relational systematicity in vision

A parallel can be drawn between this idea and the approach to systematicity used in the body of the present paper, by comparing the multiple-interpretation approach to semantics to the multiple-measurement encoding of visual objects. To realize the full extent of the analogy, one may identify the interpretation (the "semantics") of a given object u (an "expression") with the *set of measurement functions* $\{m_i\}$ responding to that object. As shown next, this representation is systematic in Hummel's (and, we believe, in Fodor's) intuitive sense. Moreover, it is both systematic and non-compositional in the formal sense of (Nerbonne, 1995) and (Lappin and Zadrozny, 1999).

Let I be the set of images (intensity functions defined over some two-dimensional "window" region $\mathcal{W}$ of R^2), and $U \subseteq I$ — the set of objects that may appear within these images. We denote two objects, $u, v \in U$, which differ only in location (i.e., $\exists \mathbf{t} \in T$, T being the set of translations acting on members of U, such that $\mathbf{t}(u) = v$) as $u \overset{T}{\leftrightarrow} v$. In what follows, a spatial reference frame encoding relative object locations will serve as the counterpart of the abstract, symbolic compositional frame played by the relation R in the propositional example (Definition 1).

Let $\mathcal{M}$ be a set of *measurement* functions, each defined over a window $W \subseteq \mathcal{W}$ and parameterized by location $\mathbf{t} \in T$, so that $m : U \times T \rightarrow R$.[14] The role of a measurement function is to provide a *perceptual symbol* (Barsalou, 1999), which stands for a particular visual event (namely, the presence in the window W of a certain pattern), and is thereby *grounded* (Harnad, 1990) in the image. As we shall see next, such measurement functions can be used to make the representation of visual objects systematic. The basic idea is to consider relations that are literally two-place (that is, are defined over two spatial locations) as the visual counterpart of the relation R from Definition 1 (as in aRb, or, equivalently, $R(a, b)$), and to construct these from localized measurements $m_i \in \mathcal{M}$.

Definition 3 (Systematic measurement space) *The set of measurement functions $\mathcal{M}$ can support effectively systematic representation if:*

M1 For the class of stimuli of interest U, any two locations can be discriminated by measurement functions belonging to a class $M \subseteq \mathcal{M}$:

$$\forall u, v \in U \text{ such that } u \overset{T}{\leftrightarrow} v, \exists m \in M \subseteq \mathcal{M} : \quad m(u;\cdot) \neq m(v;\cdot)$$

Without this condition, there would be no two-place relations, let alone systematic ones.

M2 Any two stimuli $u, v \in U$ that can be discriminated at some location t_1 can also be discriminated at any other location t_2 that is distinguishable from it in the sense of M1:

$$\exists m_i \in M : m_i(u; t_1) \neq m_i(v; t_1) \quad \Leftrightarrow \quad \exists m_j \in M : m_j(u; t_2) \neq m_j(v; t_2)$$

[14] Note that the structural description scheme such as Recognition By Components (Biederman, 1987) is subsumed by the present framework, if the measurement function is made to return a symbolic label which categorizes the object as a member of a small set of generic shapes.

In the terminology of Definition 1, this corresponds to the requirement that a and b be distinguishable in any of the two argument slots of R. Without this condition, interchanging the arguments around (as called for by the standard notion of configurational systematicity) could in principle lead to a failure of systematicity merely for the (trivial) reason of confusion between the objects that enter into the relation R.

Corollary 4 *Consider the two-place relation*

$$R\left(t_1\left(u\right);t_2\left(v\right)\right) \doteq \left(m_i\left(u;t_1\right),m_j\left(v;t_2\right)\right)$$

A measurement system that meets conditions M1 and M2 is systematic in the sense of Definition 1: its ability to deal with any $R(a,b) = aRb$ (that is, to distinguish it from some other $R(x,y)$) entails the ability to deal with $R(b,a)$ (while distinguishing it from any $R(u,v)$). To realize that, substitute in Definition 1 u at t_1 for a and v at t_2 for b, and apply Definition 3.

The immediacy of this conclusion underscores an observation we made in section 2.4 and elsewhere in this paper: systematicity is easy if dynamic binding or, equivalently, symbol manipulation, is allowed. It is the challenge of implementation in neuronal hardware that makes the modeling of systematicity interesting.

References

Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1974). *The design and analysis of computer algorithms*. Addison-Wesley, Reading, MA.

Arbib, M. A. (1979). Feature detectors, visuomotor coordination, and efferent control. In Albrecht, D., editor, *Recognition of Pattern and Form*, volume 44 of *Lecture Notes in Biomathematics*, pages 100–110. Springer, Berlin.

Bar, M. and Biederman, I. (1998). Subliminal visual priming. *Psychological Science*, 9(6):464–469.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147.

Bienenstock, E. (1996). Composition. In Aertsen, A. and Braitenberg, V., editors, *Brain Theory - Biological Basis and Computational Theory of Vision*, pages 269–300. Elsevier.

Bienenstock, E. and Geman, S. (1995). Compositionality in neural systems. In Arbib, M. A., editor, *The handbook of brain theory and neural networks*, pages 223–226. MIT Press.

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

Burl, M. C., Weber, M., and Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. 4^{th} Europ. Conf. Comput. Vision, H. Burkhardt and B. Neumann (Eds.), LNCS-Series Vol. 1406–1407, Springer-Verlag*, pages 628–641.

Cave, K. R., Pinker, S., Giorgi, L., Thomas, C. E., Heller, L. M., Wolfe, J. M., and Lin, H. (1994). The representation of location in visual images. *Cognitive Psychology*, 26:1–32.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61.

Chipere, N. (1997). Individual differences in syntactic skill. *Working Papers in English and Applied Linguistics*, 4:1–32.

Church, A. (1941). *The Calculi of Lambda Conversion*. Princeton University Press, Princeton, NJ.

Clark, A. (2000). *A theory of sentience*. Oxford University Press, Oxford.

Connor, C. E., Preddie, D. C., Gallant, J. L., and Van Essen, D. C. (1997). Spatial attention effects in macaque area V4. *J. of Neuroscience*, 17:3201–3214.

Crawford, L. E., Regier, T., and Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization. *Cognition*, 75:209–235.

Dickinson, S., Bergevin, R., Biederman, I., Eklundh, J., Munck-Fairwood, R., Jain, A., and Pentland, A. (1997). Panel report: The potential of geons for generic 3-d object recognition. *Image and Vision Computing*, 15:277–292.

Dill, M. and Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception*, 30:707–724.

Eco, U. (1976). *A theory of semiotics*. Indiana University Press, Bloomington, IN.

Edelman, S. (1994). Biological constraints and the representation of structure in vision and language. *Psycoloquy*, 5(57). http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?5.57.

Edelman, S. (1997). Computational theories of object recognition. *Trends in Cognitive Science*, 1:296–304.

Edelman, S. (1998). Representation is representation of similarity. *Behavioral and Brain Sciences*, 21:449–498.

Edelman, S. (1999). *Representation and recognition in vision*. MIT Press, Cambridge, MA.

Edelman, S. (2002). Constraints on the nature of the neural representation of the visual world. *Trends in Cognitive Sciences*, 6:125–131.

Edelman, S. and Duvdevani-Bar, S. (1997a). A model of visual recognition and categorization. *Phil. Trans. R. Soc. Lond. (B)*, 352(1358):1191–1202.

Edelman, S. and Duvdevani-Bar, S. (1997b). Similarity-based viewspace interpolation and the categorization of 3D objects. In *Proc. Similarity and Categorization Workshop*, pages 75–81, Dept. of AI, University of Edinburgh.

Edelman, S., Hiles, B. P., Yang, H., and Intrator, N. (2002). Probabilistic principles in unsupervised learning of visual structure: human data and a model. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 19–26, Cambridge, MA. MIT Press.

Edelman, S. and Intrator, N. (1997). Learning as extraction of low-dimensional representations. In Medin, D., Goldstone, R., and Schyns, P., editors, *Mechanisms of Perceptual Learning*, pages 353–380. Academic Press.

Edelman, S. and Intrator, N. (2000). (Coarse Coding of Shape Fragments) + (Retinotopy) $\approx$ Representation of Structure. *Spatial Vision*, 13:255–264.

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

Edelman, S., Intrator, N., and Jacobson, J. S. (2003). Unsupervised learning of visual structure. In Bülthoff, H. H., Wallraven, C., Lee, S.-W., and Poggio, T., editors, *Proc. 2nd Intl. Workshop on Biologically Motivated Computer Vision*, volume 2525. Springer. Lecture Notes in Computer Science, in press.

Elliffe, M. C. M., Rolls, E. T., and Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics*, 86:59–71.

Field, D. J. (1993). Scale-invariance and self-similar wavelet transforms: An analysis of natural scenes and mammalian visual systems. In Farge, M., Hunt, J., and Vassilicos, T., editors, *Wavelets, Fractals and Fourier Transforms: New Developments and new applications*, pages 151–193. Oxford University Press.

Fodor, J. and McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35:183–204.

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

Fodor, J. A. (1987). *Psychosemantics*. MIT Press, Cambridge, MA.

Fodor, J. A. (1998). *Concepts: where cognitive science went wrong*. Clarendon Press, Oxford.

Frege, G. (1891/1993). On sense and reference. In Geach, P. and Black, M., editors, *Translations from the Philosophical Writings of G. Frege*, pages 56–78. Blackwell, Oxford. Translated as "On Sense and Meaning".

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130.

Gallant, J. L., Shoup, R. E., and Mazer, J. A. (2000). A human extrastriate area functionally homologous to Macaque V4. *Neuron*, 27:227–235.

Geman, S. (1996). Minimum Description Length priors for object recognition. In *Challenging the frontiers of knowledge using statistical science (Proc. JSM'96)*.

Gilmore, G., Hersh, H., Caramazza, A., and Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception and Psychophysics*, 25:425–431.

Hadley, R. F. (1994). Systematicity revisited. *Mind and Language*, 9:431–444.

Hadley, R. F. (1997). Cognition, systematicity, and nomic necessity. *Mind and Language*, 12:137–153.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Hayward, W. G. and Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55:39–84.

Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of the National Academy of Science*, 93:623–627.

Heisele, B., Serre, T., Pontil, M., Vetter, T., and Poggio, T. (2002). Categorization by learning and combining object parts. In Becker, S., editor, *Advances in Neural Information Processing Systems 14*, pages 1239–1245. MIT Press.

Hinton, G. E. (1984). Distributed representations. Technical Report CMU-CS 84-157, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Hummel, J. E. (2000). Where view-based theories of human object recognition break down: the role of structure in human shape perception. In Dietrich, E. and Markman, A., editors, *Cognitive Dynamics: conceptual change in humans and machines*, chapter 7. Erlbaum, Hillsdale, NJ.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, 8:489–517.

Hummel, J. E. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99:480–517.

Huttenlocher, J., Hedges, L., and Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological Review*, 98:352–376.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Johnston, J. C. and Pashler, H. (1990). Close binding of identity and location in visual feature perception. *Journal of Experimental Psychology: Human Perception and Performance*, 16:843–856.

Jolicoeur, P., Gluck, M., and Kosslyn, S. M. (1984). Pictures and names: making the connection. *Cognitive Psychology*, 16:243–275.

Kahneman, D., Treisman, A., and Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, 24:175–219.

Katz, J. J. and Fodor, J. (1963). The structure of a semantic theory. *Language*, 39:17–210.

Kaye, L. J. (1995). The languages of thought. *Philosophy of Science*, 62:92–110.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, 71:856–867.

Lakoff, G. (1971). On generative semantics. In Steinberg, D. D. and Jakobowitz, L. A., editors, *Semantics*, pages 232–296. Cambridge University Press.

Langacker, R. W. (1990). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Mouton de Gruyter, Berlin.

Lappin, S. and Zadrozny, W. (1999). Compositionality, synonymy, and the systematic representation of meaning. unpublished manuscript, King's College London and IBM T. J. Watson Research Center, Hawthorne, NY.

Lewicki, M. S. and Sejnowski, T. J. (1998). Learning nonlinear overcomplete representations for efficient coding. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20:1015–1036.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape recognition in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563.

Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19:577–621.

Love, B. C., Rouder, J. N., and Wisniewski, E. J. (1999). A structural account of global and local processing. *Cognitive Psychology*, 38:291–316.

Lund, J. S., Yoshita, S., and Levitt, J. B. (1993). Comparison of intrinsic connections in different areas of macaque cerebral cortex. *Cerebral Cortex*, 3:148–162.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 4:414–417.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9:353–383.

Nazir, T. and O'Regan, J. K. (1990). Some results on translation invariance in the human visual system. *Spatial vision*, 5:81–100.

Nerbonne, J. (1995). Computational semantics – linguistics and processing. In Lappin, S., editor, *Handbook of Contemporary Semantic Theory*, pages 461–484. Blackwell, London.

Niklasson, L. and Boden, M. (1997). Representing structure and structured representations in connectionist networks. In Browne, A., editor, *Neural Network Perspectives on Cognition and Adaptive Robotics*, pages 20–50. Institute of Physics Press, Bristol, UK.

Niklasson, L. and van Gelder, T. (1994). Can connectionist models exhibit non-classical structure sensitivity? In Ram, A. and Eiselt, K., editors, *Proc. Sixteenth Annual Conference of the Cognitive Science Society*, pages 664 – 669, Hillsdale, NJ. Erlbaum.

O'Brien, G. and Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, 22:127–148.

Op de Beeck, H. and Vogels, R. (2000). Spatial sensitivity of Macaque inferior temporal neurons. *J. Comparative Neurology*, 426:505–518.

O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:883–917.

O'Reilly, R. C. and Busby, R. S. (2002). Generalizable relational binding from coarse-coded distributed representations. In Becker, S., editor, *Advances in Neural Information Processing Systems 14*, pages 75–82–. MIT Press.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Poggio, T. and Hurlbert, A. (1994). Observations on cortical mechanisms for object recognition and learning. In Koch, C. and Davis, J., editors, *Large Scale Neuronal Theories of the Brain*, pages 153–182. MIT Press, Cambridge, MA.

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

Prinz, J. (1994). Shared beliefs, infralingual thought, and nativism. *Noesis*, online(#004986). http://noesis.evansville.edu.

Rainer, G., Asaad, W., and Miller, E. K. (1998). Memory fields of neurons in the primate prefrontal cortex. *Proceedings of the National Academy of Science*, 95:15008–15013.

Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276:821–824.

Riesenhuber, M. and Dayan, P. (1997). Neural models for the part-whole hierarchies. In Jordan, M., editor, *Advances in Neural Information Processing*, volume 9, pages 17–23. MIT Press.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025.

Rock, I. (1973). *Orientation and form*. MIT Press, Cambridge, MA.

Rolls, E. T. (1996). Visual processing in the temporal lobe for invariant object recognition. In Torre, V. and Conti, T., editors, *Neurobiology*, pages 325–353. Plenum Press, New York.

Rolls, E. T. and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. of Neurophysiology*, 73:713–726.

Roskies, A. L. (1999). The binding problem (review introduction). *Neuron*, 24:7–9.

Saiki, J. and Hummel, J. E. (1996). Attribute conjunctions and the part configuration advantage in object category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:1002–1019.

Sali, E. and Ullman, S. (1999). Detecting object classes by the detection of overlapping 2-D fragments. In *Proc. 10th British Machine Vision Conference*, volume 1, pages 203–213.

Salinas, E. and Abbott, L. F. (1997). Invariant visual responses from attentional gain fields. *J. of Neurophysiology*, 77:3267–3272.

Sanocki, T. (1993). Time course of object identification: evidence for a global-to-local contingency. *Journal of Experimental Psychology: Human Perception and Performance*, 19:878–898.

Schlingensiepen, K.-H., Campbell, F. W., Legge, G. E., and Walker, T. D. (1986). The importance of eye movements in the analysis of simple patterns. *Vision Research*, 26:1111–1117.

Shapiro, K. L. and Loughlin, C. (1993). The locus of inhibition in the priming of static objects: object token versus location. *Journal of Experimental Psychology: Human Perception and Performance*, 19:352–363.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–397.

Smith, B. (2001). Fiat objects. *Topoi*, 20:131–148.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46:159–216.

Solomon, K. O. and Barsalou, L. W. (2001). Representing properties locally. *Cognitive Psychology*, 43:129–169.

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

Strang, G. (1989). Wavelets and dilation equations: a brief introduction. *SIAM Review*, 31:614–627.

Talmy, L. (1983). How language structures space. In H. L. Pick, J. and Acredolo, L. P., editors, *Spatial Orientation: Theory, Research, and Application*, pages 225–282. Plenum Press, New York.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophysiol.*, 66:170–189.

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, 9:483–484.

Touretzky, D. S. (1989). Connectionism and compositional semantics. Technical Report CMU-CS 89-147, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Treisman, A. (1992). Perceiving and re-perceiving objects. *American Psychologist*, 47:862–875.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, 6:171–178.

Treisman, A. M. and Kanwisher, N. G. (1998). Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 8:218–226.

Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4:832–838.

Ullman, S. (1989). Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, 67:21–44.

Ungerleider, L. and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4:157–165.

van Gelder, T. (1990). Compositionality: A connectionist variation on a theme. *Cognitive Science*, 14:355–384.

Venturino, M. and Gagnon, D. (1992). Information tradeoffs in complex stimulus structure: local and global levels in naturalistic scenes. *Perception and Psychophysics*, 52:425–436.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur. J. Neurosci.*, 11:1239–1255.

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5:520–526.

Wallach, H. and Austin-Adams, P. (1954). Recognition and the localization of visual traces. *American Journal of Psychology*, 67:338–340.

Wang, G., Tanifuji, M., and Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neurosci. Res.*, 32:33–46.

Wang, Y., Fujita, I., and Murayama, Y. (2000). Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. *Nature Neuroscience*, 3:807–813.

Ward, R., Danziger, S., Owen, V., and Rafal, R. (2002). Deficits in spatial coding and feature binding following damage to spatiotopic maps in the human pulvinar. *Nature Neuroscience*, 5:99–101.

Wittgenstein, L. (1961). *Tractatus Logico-philosophicus*. Routledge, London. trans. D. F. Pears and B. F. McGuinness.

Wolfe, J. M. and Bennett, S. C. (1997). Preattentive object files: Shapeless bundles of basic features. *Vision Research*, 37:25–43.

Zadrozny, W. (1994). From compositional to systematic semantics. *Linguistics and philosophy*, 17:329–342.

Zadrozny, W. (1999). Minimum description length and compositionality. In Bunt, H. and Muskens, R., editors, *Computing Meaning*, volume 1, pages 113–128. Kluwer.

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