ARTICLE Communicated by Shimon Edelman

Minimizing Binding Errors Using Learned Conjunctive Features

Bartlett W. Mel
Department of Biomedical Engineering, University of Southern California, Los Angeles, CA 90089, U.S.A.

József Fiser
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627-0268, U.S.A.

We have studied some of the design trade-offs governing visual representations based on spatially invariant conjunctive feature detectors, with an emphasis on the susceptibility of such systems to false-positive recognition errors—Malsburg's classical binding problem. We begin by deriving an analytical model that makes explicit how recognition performance is affected by the number of objects that must be distinguished, the number of features included in the representation, the complexity of individual objects, and the clutter load, that is, the amount of visual material in the field of view in which multiple objects must be simultaneously recognized, independent of pose, and without explicit segmentation. Using the domain of text to model object recognition in cluttered scenes, we show that with corrections for the nonuniform probability and nonindependence of text features, the analytical model achieves good fits to measured recognition rates in simulations involving a wide range of clutter loads, word sizes, and feature counts. We then introduce a greedy algorithm for feature learning, derived from the analytical model, which grows a representation by choosing those conjunctive features that are most likely to distinguish objects from the cluttered backgrounds in which they are embedded. We show that the representations produced by this algorithm are compact, decorrelated, and heavily weighted toward features of low conjunctive order. Our results provide a more quantitative basis for understanding when spatially invariant conjunctive features can support unambiguous perception in multiobject scenes, and lead to several insights regarding the properties of visual representations optimized for specific recognition tasks.

1 Introduction

The problem of classifying objects in visual scenes has remained a scientific and a technical holy grail for several decades. In spite of intensive research in

the field of computer vision and a large body of empirical data from the fields of visual psychophysics and neurophysiology, the recognition competence of a two-year-old human child remains unexplained as a theoretical matter and well beyond the technical state of the art. This fact is surprising given that (1) the remarkable speed of recognition in the primate brain allows for only very brief processing times at each stage in a pipeline containing only a few stages (Potter, 1976; Oram & Perrett, 1992; Heller, Hertz, Kjær, & Richmond, 1995; Thorpe, Fize, & Marlot, 1996; Fize, Boulanouar, Ranjeva, Fabre-Thorpe, & Thorpe, 1998), (2) the computations that can be performed within each neural processing stage are strongly constrained by the structure of the underlying neural tissue, about which a great deal is known (Hubel & Wiesel, 1968; Szentagothai, 1977; Jones, 1981; Gilbert, 1983; Van Essen, 1985; Douglas & Martin, 1998), (3) the response properties of neurons in each of the relevant brain areas have been well studied, and evolve from stage to stage in systematic ways (Oram & Perrett, 1994; Logothetis & Sheinberg, 1996; Tanaka, 1996), and (4) computer systems may already be powerful enough to emulate the functions of these neural processing stages, were it only known what exactly to do.

A number of neuromorphic approaches to visual recognition have been proposed over the years (Pitts & McCullough, 1947; Fukushima, Miyake, & Ito, 1983; Sandon & Urh, 1988; Zemel, Mozer, & Hinton, 1990; Le Cun et al., 1990; Mozer, 1991; Swain & Ballard, 1991; Hummel & Biederman, 1992; Lades et al., 1993; Califano & Mohan, 1994; Schiele & Crowley, 1996; Mel, 1997; Weng, Ahuja, & Huang, 1997; Lang & Seitz, 1997; Wallis & Rolls, 1997; Edelman & Duvdevani-Bar, 1997). Many of these neurally inspired systems involve constructing banks of feature detectors, often called receptive fields (RFs), each of which is sensitive to some localized spatial configuration of image cues but invariant to one or more spatial transformations of its preferred stimulus—typically including invariance to translation, rotation, scale, or spatial distortion, as specified by the task.[1] Since the goal to extract useful invariant features is common to most conventional approaches to computer vision as well, the "neuralness" of a recognition system lies primarily in its emphasis on feedforward, hierarchical construction of the invariant features, where the computations are usually restricted to simple spatial conjunction and disjunction operations (e.g., Fukushima et al., 1983), and/or specify the inclusion of a relatively large number of invariant features in the visual representation (see Mozer, 1991; Califano & Mohan, 1994; Mel, 1997)—anywhere from hundreds to millions. Neural approaches typically also involve some form of learning, whether supervised by category labels at the output layer (Le Cun et al., 1990), supervised directly within

[1] A number of the above-mentioned systems have emphasized other mechanisms instead of or in addition to straightforward image filtering operations (Zemel et al., 1990; Mozer, 1991; Hummel & Biederman, 1992; Lades et al., 1993; Califano & Mohan, 1994; Edelman & Duvdevani-Bar, 1997).

intermediate network layers (Fukushima et al., 1983), or involving purely unsupervised learning principles that home in on features that occur frequently in target objects (Fukushima et al., 1983; Zemel et al., 1990; Wallis & Rolls, 1997; Weng et al., 1997). While architectures of this general type have performed well in a variety of difficult, though limited, recognition problems, it has yet to be proved that a few stages of simple feedforward filtering operations can explain the remarkable recognition and classification capacities of the human visual system, which is commonly confronted with scenes acquired from varying viewpoints, containing multiple objects drawn from thousands of categories, and appearing in an infinite panoply of spatial configurations. In fact, there are reasons for pessimism.

1.1 The Binding Problem. One of the most persistent worries concerning this class of architecture is due to Malsburg (1981), who noted the potential for ambiguity in visual representations constructed from spatially invariant filters. A spatial binding problem arises, for example, when each detector in the visual representation reports only the *presence* of some elemental object attribute but not its spatial location (or, more generally, its pose). Under these unfavorable conditions, an object is hallucinated (i.e., detected when not actually present) whenever all of its elemental features are present in the visual field, even when embedded piecemeal in improper configurations within a scattered coalition of distractor objects (see Figure 1A). This type of failure mode for a visual representation emanates from the dual pressures to cope with uncontrolled variation in object pose, which forces the use of detectors with excessive spatial invariances, and the necessity to process multiple objects simultaneously, which overstimulate the visual representation and increase the probability of ambiguous perception.

One approach to the binding problem is to build a separate, full-order conjunctive detector for every possible view (e.g., spatial pose, state of occlusion, distortion, degradation) of every possible object, and then for each object provide a huge disjunction that pools over all of the views of that object.

Malsburg (1981) pointed out that while the binding problem could thus be solved, the remedy is a false one since it leads to a combinatorial explosion in the number of needed visual processing operations (see Figure 1B). Other courses of action include (1) strategies for image preprocessing designed to segment scenes into regions containing individual objects, thus reducing the clutter load confronting the visual representation, and (2) explicit normalization procedures to reduce pose uncertainty (e.g., centering, scaling, warping), thereby reducing the invariance load that must be sustained by the individual receptive fields. Both strategies reduce the probability that any given object's feature set will be activated by a spurious collection of features drawn from other objects.

1.2 Wickelsystems. Preprocessing strategies aside, various workers have explored the middle ground between "binding breakdown" and "com-

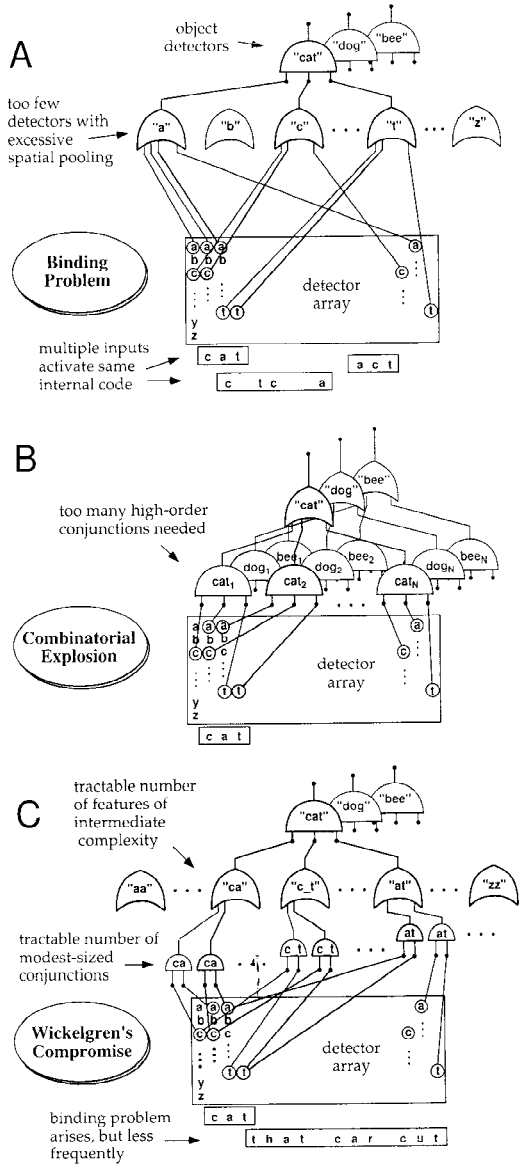


Figure 1: (A) A binding problem occurs when an object's representation can be falsely activated by multiple (or inappropriate) other objects. (B) A combinatorial explosion arises when all possible views of all possible objects are represented by explicit conjunctions. (C) Compromise representation containing an intermediate number of receptive fields that bind image features to intermediate order.

binatorial catastrophe," where a visual representation contains an intermediate number of detectors, each of which binds together an intermediate-sized subset of feature elements in their proper relations before integrating out the irrelevant spatial transformations (see Mozer, 1991, for a thoughtful advocacy of this position). Under this compromise, each detector becomes a spatially invariant template for a specific minipattern, which is more complex than an elemental feature, such as an oriented edge, but simpler than a full-blown view of an object (see Figure 1C). A set of spatially invariant features of intermediate complexity can be likened to a box of jumbled jigsaw puzzle pieces: although the final shape of the composite object—the puzzle—is not explicitly contained in this peculiar "representation," it is nonetheless unambiguously specified: the collection of parts (features) can interlock in only one way.

An important early version of this idea is due to Wickelgren (1969), who proposed a scheme for representing the phonological structure of spoken language involving units sensitive to contiguous triples of phonological features but not to the absolute position of the tuples in the speech stream. Though not phrased in this language, "Wickelfeatures" were essentially translation-invariant detectors of phonological minipatterns of intermediate complexity, analyzing the one-dimensional speech signal. Representational schemes inspired by Wickelgren's proposal have since cropped up in other influential models of language processing (McClelland & Rumelhart, 1981; Mozer, 1991).

The use of invariant conjunctive features schemes to address representational problems in fields other than vision highlights the fact that the binding problem is not an inherently visuospatial one. Rather, it can arise whenever the set of features representing an object (visual, auditory, olfactory, etc.) can be fully, and thus inappropriately, activated by input "scenes" that do not contain the object. In general, a binding problem has the potential to exist whenever a representation lacks features that conjoin an object's elemental attributes (e.g., parts, spatial relations, surface properties) to sufficiently high conjunctive order. On the solution side, binding ambiguity can be mitigated as in the scenario of Figure 1 by building conjunctive features that are increasingly object specific, until the probability of false-positive activation for any object is driven below an acceptable threshold. Given that this conjunctive order-boosting prescription to reduce ambiguity is algorithmically trivial, it shifts emphasis away from the question as to how to eliminate binding ambiguity, toward the question as to whether, for a given recognition problem, ambiguity can be eliminated at an acceptable cost.

1.3 A Missing Theory. In spite of the importance of spatially invariant conjunctive visual representations as models of neural function, and the demonstrated practical successes of computer vision systems constructed along these lines, many of the design trade-offs that govern the performance of visual "Wickelsystems" in response to cluttered input images have re-

mained poorly understood. For example, it is unknown in general how recognition performance depends on (1) the number of object categories that must be distinguished, (2) the similarity structure of the object category distribution (i.e., whether object categories are intrinsically very similar or very different from each other), (3) the featural complexity of individual objects, (4) the number and conjunctive order of features included in the representation, (5) the clutter load (i.e., the amount of visual material in the field of view from which multiple objects must be recognized without explicit segmentation), and (6) the invariance load (i.e., the set of spatial transformations that do not affect the identities of objects, and that must be ignored by each individual detector). In a previous analysis that touched on some of these issues, Califano and Mohan (1994) showed that large, parameterized families of complex invariant features (e.g., containing 10^6 elements or more) could indeed support recognition of multiple objects in complex scenes without prior segmentation or discrimination among a large number of highly similar objects in a spatially invariant fashion. However, these authors were primarily concerned with the mathematical advantages of large versus small feature sets in a generic sense, and thus did not test their analytical predictions using a large object database and varying object complexity, category structure, clutter load, and other factors.

Beyond our lack of understanding regarding the trade-offs influencing system performance, the issue as to which invariant conjunctive features and of what complexity should be incorporated into a visual representation, through learning, has also not been well worked out as a conceptual matter. Previous approaches to feature learning in neuromorphic visual systems have invoked (1) generic gradient-based supervised learning principles (e.g., Le Cun et al., 1990) that offer no direct insight into the qualities of a good representation, (2) or straightforward unsupervised learning principles (e.g., Wallis & Rolls, 1997), which tend to discover frequently occurring features or principal components rather than those needed to distinguish objects from cluttered backgrounds on a task-specific basis. Network approaches to learning have also typically operated within a fixed network architecture, which largely predefines and homogenizes the complexity of top-level features to be used for recognition, though optimal representations may actually require features drawn from a range of complexities, and could vary in their composition on a per-object basis. All this suggests that further work on the principles of feature learning is needed.

In this article, we describe our work to test a simple analytical model that captures several trade-offs governing the performance of visual recognition systems based on spatially invariant conjunctive features. In addition, we introduce a supervised greedy algorithm for feature learning that grows a visual representation in such a way as to minimize false-positive recognition errors. Finally, we consider some of the surprising properties of "good" representations and the implications of our results for more realistic visual recognition problems.

2 Methods

2.1 *Text World* as a Model for Object Recognition. The studies described here were carried out in *text world*, a domain with many of the complexities of vision in general (see Figure 2): target objects are numerous (more than 40,000 words in the database), are highly nonuniform in their prior probabilities (relative probabilities range from 1 to 400,000), are constructed from a set of relatively few underlying parts (26 letters), and individually can contain from 1 to 20 parts. Furthermore, the parts from which words are built are highly nonuniform in their relative frequencies and contain strong statistical dependencies. Finally, word objects occur in highly cluttered visual environments—embedded in input arrays in which many other words are simultaneously present—but must nonetheless be recognized regardless of position in the visual field.

Although *text world* lacks several additional complexities characteristic of real object vision (see section 6), it nonetheless provides a rich but tractable domain in which to quantify binding trade-offs and to facilitate the testing of analytical predictions and the development of feature learning algorithms. An important caveat, however, is that the use of text as a substrate for developing and testing our analytical model does not imply that our conclusions bear directly on any aspect of human *reading* behavior.

Two databases were used in the studies. The word database contained 44,414 entries, representing all lowercase punctuation-free words and their relative frequencies found in 5 million words in the *Wall Street Journal* (WSJ) (available online from the Association for Computational Linguistics at http://morph.ldc.upenn.edu/). The English text database consisted of approximately 1 million words drawn from a variety of sources (Kučera & Francis, 1967).

2.2 Recognition Using *n*-Grams. A natural class of visual features in *text world* is the position-invariant *n*-gram, defined here as a binary detector that responds when a specific spatial configuration of *n* letters is found anywhere (one or more times) in the input array.[2] The value of *n* is termed the conjunctive or binding order of the *n*-gram, and the diameter of the *n*-gram is the maximum span between characters specified within it. For example, th**_ is a 3-gram of diameter 5 since it specifies the relative locations of three characters (including the space character _ but not wild card characters *), and spans a field five characters in width. The concept of an *n*-gram is used also in computational linguistics, though usually referring to *n*-tuples of consecutive words (Charniak, 1993).

[2] An *n*-gram could also be multivalued, that is, respond in a graded fashion depending on the number of occurrences of the target feature in the input; the binary version was used here to facilitate analysis.

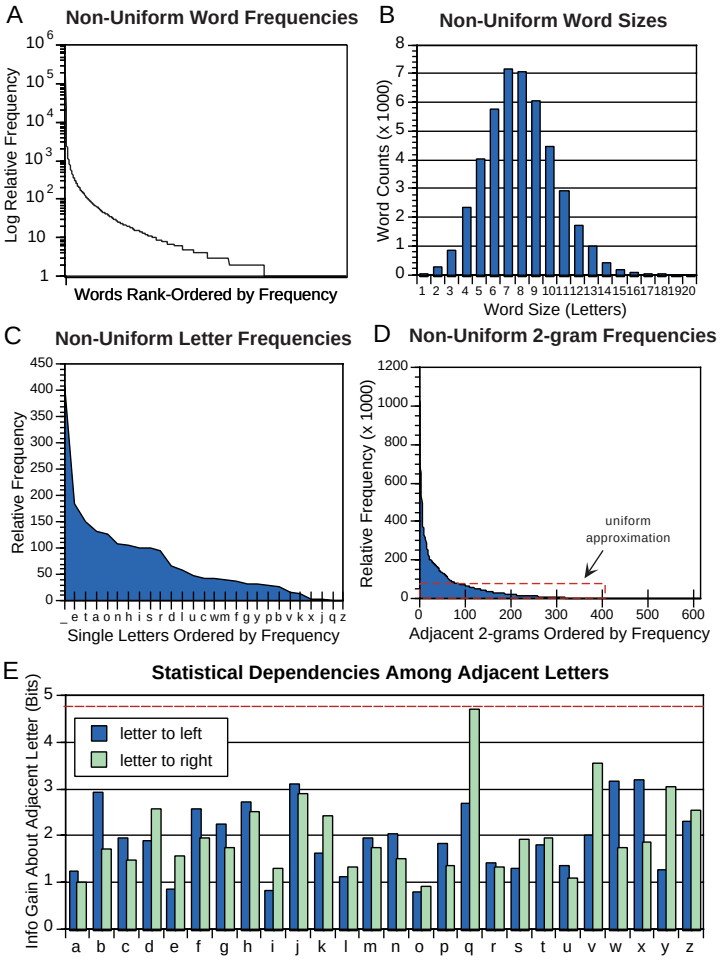


Figure 2: *Text world* was used as a surrogate for visual recognition in multiobject scenes. (A) Word frequencies in the 5-million-word *Wall Street Journal* (WSJ) corpus are highly nonuniform, following an approximately exponential decay according to Zipf's law. (B) Words range from 1 to 20 characters in length; 7-letter words are most numerous. (C,D) Relative frequencies of individual letters or adjacent letter pairs are highly nonuniform. Ordinate values represent the number of times the feature was found in the WSJ corpus. The dashed rectangle in *D* represents simplifying assumption regarding the 2-gram frequency distribution used for quantitative predictions below. (E) The parts of words show strong statistical dependencies. Columns show information gain about identity of the adjacent letter (or space character) to left or right, given the presence of letter indicated on the x-axis. The dashed horizontal line indicates perfect knowledge—$\log_2(27) = 4.75$ bits.

In relation to conventional visual domains, *n*-grams may be viewed as complex visual feature detectors, each of which responds to a specific conjunction of subparts in a specific spatial configuration, but with built-in invariance to a set of spatial transformations predefined by the recognition task. In relation to a shape-based visual domain, a 2-gram is analogous to a corner detector that signals the co-occurrence of a vertical and a horizontal edge, with appropriate spatial offsets, anywhere within an extended region of the image.

Let $\mathcal{R}$ be a representation containing a set of position-invariant *n*-grams analyzing the input array and $\mathcal{W}$ be a set of word detectors analyzing $\mathcal{R}$, with one detector for every entry in the word database. Each word detector receives input from some or all of the *n*-gram units in $\mathcal{R}$ that are contained in the respective word. However, not all of a word's *n*-grams are necessarily contained in $\mathcal{R}$, and not all of the word's *n*-grams that *are* contained in $\mathcal{R}$ necessarily provide input to the word's detector. For example, the representation might contain four spatially invariant features $\mathcal{R} = \{a, c, o, s\}$, while the detector for the word *cows* receives input from features *c* and *s*, but not from the *o* feature, which is present in $\mathcal{R}$, or from the *w* feature, which is not contained in $\mathcal{R}$. Word detectors are binary-valued conjunctions, responding 1 when all of their inputs are active and 0 otherwise. The collection of *n*-grams feeding a word detector is hereafter referred to as the word's *minimal feature set*.

A correct recognition is scored when an input array of characters is presented, and each word detector is activated if and only if its associated word is present in the input. Any word detector activated when its corresponding word is not actually present in the input array is counted as a hallucination, and the recognition trial is scored as an error. Recognition errors for a particular representation $\mathcal{R}$ are reported as the percentage of input arrays drawn randomly from the English text database that generated one or more word hallucinations.

3 Results

3.1 Word-Word Collisions in a Simple 1-Gram Representation. We begin by considering the case without visual clutter and ask how often pairs of isolated words collide with each other within a binary 1-gram representation, where each word is represented as an unordered set of individual letters (with repeats ignored). This is the situation in which the binding problem is most evident, as schematized in Figure 1A. Results for the approximately 2 billion distinct pairs of words in the database are summarized in Table 1. About half of the 44,414 words contained in the database collide with at least one other word (e.g., `cerebrum` ≡ `cucumber`, `calculation` ≡ `unconstitutional`). The largest cohort of 1-gram-identical words contained 28 members.

Table 1: Performance of a Binary 1-Gram Representation in Word-Word Comparisons.

Quantity	Value	Comment
Number of 1-grams in $\mathcal{R}$	27	a, b, . . . , z, _
Number of distinct words	44,414	From 5 million words in the WSJ
Word-word comparisons	∼2 billion	

Number of ambiguous words	24,488	`analogies` ≡ `gasoline`
		`suction` ≡ `continuous`
		`scientists` ≡ `nicest`

Largest self-similar cohort	28 words	`stare, arrest, tears, restates, reassert, rarest, easter` ...
Baseline entropy in word-frequency distribution	9.76 bits	$< \log_2(44,414) = 15.4$ bits
Residual uncertainty about identity of a randomly drawn word, knowing its 1-grams	1.4 bits	Narrows field to ∼ 3 possible words

As an aside, we noted that if the two words compared were required also to contain the same number of letters, the number of unambiguously represented words fell to 11,377, and if the words were compared within a multivalued 1-gram representation (where words matched only if they contained the identical multiset of letters, i.e., in which repeats were taken into account), the ambiguous word count fell further to 5,483, or about 12% of English words according to this sample. The largest self-similar cohort in this case contained only 5 words (`traces`, `reacts`, `crates`, `caters`, `caster`).

Although a 1-gram representation contains no explicit part-part relations and fails to pin down uniquely the identity of more than half of the English words in the database, it nonetheless provides most of the needed information to distinguish individually presented words. The baseline entropy in the WSJ word-frequency distribution is 9.76 bits, significantly less than the $15.4 = \log_2(44{,}414)$ bits for an equiprobable word set. The average residual uncertainty about the identity of a word drawn from the WSJ word frequency distribution, given its 1-gram representation, is only 1.4 bits, meaning that for individually presented words, the 1-gram representation narrows the set of possibilities to about three words on average.

3.2 Word-Word Collisions in an Adjancent 2-Gram Representation. Since many word objects have identical 1-gram representations even when full multisets of letters are considered, we examined the collision rate when an adjacent binary-valued 2-gram representation was used (i.e., coding the

Table 2: Performance of a Binary Adjacent 2-Gram Representation in Word-Word Comparisons.

Quantity	Value	Comment
Number of adjacent 2-grams	729	[aa], [ab], . . . , [a_], . . . , [_z], [__]
Number of words	44,414	
Word-word comparisons	∼2 billion	
Number of ambiguous words	57	`ohhh, ahhh, shhh, hmmm, whoosh, whirr,...`
Largest self-similar cohort	5 words	`ohhh` ≡ `ohhhh` ≡ `ohhhhh` ≡ `ohhhhhh` ≡ `ohhhhhhh`

Ambiguous word pairs that are linguistically distinct	4 pairs	`asses` ≡ `assess`
		`possess` ≡ `possesses`
		`seamstress` ≡ `seamstresses`
		`intended` ≡ `indented`

Words with identical adjacent 2-gram multisets	2 words	`intended` ≡ `indented`

set of all adjacent letter pairs, including the space character). This representation is significant in that adjacent 2-grams are the simplest features that explicitly encode spatial relations between parts. The results of this test are summarized in Table 2. There were only 46 word-pair collisions in this case, comprising 57 distinct words. Forty-two of the 46 problematic word pairs differed only in the length of a string of repeated letters (e.g., `ahhh` ≡ `ahhhh` ≡ `ahhhhh` . . . , similarly for `ohhh, ummm, zzz, wheee, whirrr`) or into sets representing varying spellings or misspellings of the same word also involving repeated letters (e.g., `tepees` ≡ `teepees`, `llama` ≡ `lllama`). Only four pairs of distinct words collided that had different morphologies in the linguistic sense (see Table 2). When words were constrained to be the same length or to contain the identical multiset of adjacent 2-grams, only a single collision persisted: `intended` ≡ `indented`.

At the level of word-word comparisons, therefore, the binding of adjacent letter pairs is sufficient practically to eliminate the ambiguity in a large database of English words. We noted empirically that the inclusion of fewer than 20 additional nonadjacent 2-grams to the representation entirely eliminated the word-word ambiguity in this database, without recourse to multiset comparisons or word-length constraints.

4 An Analytical Model

The results of these word-word comparisons suggest that under some circumstances, a modest number of low-order *n*-grams can disambiguate a

large, complex object set. However, one of the most dire challenges to a spatially invariant feature representation, object clutter, remains to be dealt with.

To this end, we developed a simple probabilistic model to predict recognition performance when a cluttered input array containing multiple objects is presented to a bank of feature detectors. To permit analysis of this problem, we make two simplifying assumptions. First, we assume that the features contained in $\mathcal{R}$ are statistically independent. This assumption is false for English n-grams: for example, [*th*] predicts that [*he*] will follow about one in three times, whereas the baseline rate for [*he*] is only 0.5%. Second, we assume that all the features in $\mathcal{R}$ are activated with equal probability. This is also false in English: for example *ed* occurs approximately 90,000 times more often than [*mj*] (as in `ramjet`).

Proceeding nonetheless under these two assumptions, it is straightforward to calculate the probability that no object in the database is hallucinated in response to a randomly drawn input image. We first calculate the probability o that all of the features feeding a given object detector are activated by an arbitrary input image:

$$o = \frac{\binom{d-w}{c-w}}{\binom{d}{c}} = \frac{(d-w)!\,c!}{(c-w)!\,d!} \approx \left(\frac{c}{d}\right)^w, \tag{4.1}$$

where d is the total number of feature detectors in $\mathcal{R}$, w is the size of the object's minimal feature set, i.e., the number of features required by the target object (word) detector, and c is the number of features activated by a "cluttered" multiobject input image. The left-most combinatorial ratio term in equation 4.1 counts the number of ways of choosing c from the set of d detectors, including all w belonging to the target object, as a fraction of the total number of ways of choosing c of the d detectors in any fashion. The approximation in equation 4.1 holds when c and d are large compared to w, as is generally the case. Equation 4.1 is, incidentally, the value of a "hypergeometric" random variable in the special case where all of the target elements are chosen; this distribution is used, for example, to calculate the probability of picking a winning lottery ticket.

The probability that an object is hallucinated (i.e., perceived but not actually present) is given by

$$h_i = \frac{o_i - q_i}{1 - q_i}, \tag{4.2}$$

where q_i is the probability that the object does in fact appear in the input image, which by definition means it cannot be hallucinated. According to Zipf's law, however, which tells us that most objects in the world are quite uncommon (see Figure 2A), we may in some cases make the further simpli-

fication that $q_i \ll o_i$, giving $h_i \approx o_i$. We adopt this assumption in the following. (This simplification would result also from the assumption that inputs consist of object-like noise structured to trigger false-positive perceptions.)

We may now write the probability of veridical perception—the probability that no object in the database is hallucinated,

$$p_v = (1-h)^N \approx \left[1 - \left(\frac{c}{d}\right)^w\right]^N, \tag{4.3}$$

where N is the number of objects in the database. This expression assumes a homogeneous object database, where all N objects require exactly w features in $\mathcal{R}$. Given the independence assumption, the expression for a heterogeneous object database consists of a product of object-specific probabilities $p_v = \prod_i (1 - (\frac{c}{d})^{w_i})$, where w_i is the number of features required by object i. Note that objects (words) containing the same number of parts (letters) do not necessarily activate the same number of features (n-grams) in $\mathcal{R}$, depending on both the object and the composition of $\mathcal{R}$.

From an expansion to first order of equation 4.3 around $(c/d)^w = 0$, which gives $p_v = 1 - N(c/d)^w$, it may be seen that perception is veridical (i.e., $p_v \approx 1$) when $(c/d)^w \ll 1/N$. Thus, recognition errors are expected to be small when (1) a cluttered input scene activates only a small fraction of the detectors in $\mathcal{R}$, (2) individual objects depend on as many detectors in $\mathcal{R}$ as possible, and (3) the number of objects to be discriminated is not too large. The first two effects generate pressure to grow large representations, since if d is scaled up and c and w scale with it (as would be expected for statistically independent features), recognition performance improves rapidly. Conveniently, the size of the representation can always be increased by including new features of higher binding order or larger diameters, or both.

It is also interesting to note that item 1 in the previous paragraph appears to militate for a sparse visual representation, while equation 4.2 militates for a dense representation. This apparent contradiction reflects the fundamental trade-off between the need for individual objects to have large minimal feature sets in order to ward off hallucinations and the need for input scenes containing many objects to activate as few features as possible, again in order to minimize hallucinations. This trade-off is not captured within the standard conception of a feature space, since the notion of "similarity as distance" in a feature space does not naturally extend to the situation in which inputs represent multiobject scenes and require multiple output labels.

4.1 Fitting the Model to Recognition Performance in *Text World*. We ran a series of experiments to test the analytical model. Input arrays varying in width from 10 to 50 characters were drawn at random from the text database and presented to $\mathcal{R}$. The representation contained either all adjacent 2-grams, called $\mathcal{R}_{\{2\}}$ to signify a diameter of 2, or the set of all 2-grams of diameter 2 or 3, called $\mathcal{R}_{\{2,3\}}$, where $\mathcal{R}_{\{2,3\}}$ was twice the size of $\mathcal{R}_{\{2\}}$.

In individual runs, the word database was restricted to words of a single length—two-letter words ($N = 253$), five-letter words ($N = 4024$), or seven-letter words ($N = 7159$)—in order for the value of w to be kept relatively constant within a trial.[3] For each word size and input size and for each of the two representations, average values for w and c were measured from samples drawn from the word or text database. Not surprisingly, given the assumption violations noted above (feature independence and equiprobability), predicting recognition performance using these measured average values led to gross underestimates of the hallucination errors actually generated by the representation.

Two kinds of corrections to the model were made. The first correction involved estimating a better value for d, the number of n-grams contained in the representation, since some of the detectors nominally present in $\mathcal{R}_{\{2\}}$ and $\mathcal{R}_{\{2,3\}}$ were infrequently, if ever, actually activated. The value for d was thus down-corrected to reflect the number of detectors that would account for more than 99.9% of the probability density over the detector set: for $\mathcal{R}_{\{2\}}$, the representation size dropped from $d = 729$ to $d' = 409$, while for $\mathcal{R}_{\{2,3\}}$, d was cut from 1,458 to 948.

The second correction was based on the fact that the n-grams activated by English words and phrases are highly statistically redundant, meaning that a set of d' n-grams has many fewer than d' underlying degrees of freedom. We therefore introduced a redundancy factor r, which allowed us to "pretend" that the representation contained only d'/r virtual n-grams, and a word or an input image activated only w/r or c/r virtual n-grams in the representation, respectively. (Note that since c and d appear only as a ratio in the approximation of equation 4.3, their r-corrections cancel out.) By systematically adjusting the redundancy factor depending on test condition (ranging from 1.60 to 3.45), qualitatively good fits to the empirical data could be generated, as shown in Figure 3. The redundancy factor was generally larger for long words relative to short words, reflecting the fact that long English words have more internal predictability (i.e., fill a smaller proportion of the space of long strings than short words fill in the space of short strings). The r-factor was also larger for $\mathcal{R}_{\{2,3\}}$ than for $\mathcal{R}_{\{2\}}$, reflecting the fact that the increased feature overlap due to the inclusion of larger feature diameters in $\mathcal{R}_{\{2,3\}}$ leads to stronger interfeature correlations, and hence fewer underlying degrees of freedom. Thus, with the inclusion of correction factors for nonuniform activation levels and nonindependence of the features in $\mathcal{R}$, the remarkably simple relation in equation 4.3 captures the main trends that describe recognition performance in scenarios with database sizes ranging from 253 to 7159 words, words containing between

[3] Since $\mathcal{R}_{\{2\}}$ and $\mathcal{R}_{\{2,3\}}$ contained all n-grams of their respective sizes, the value of w could vary only from word to word to the extent that words contained one or more repeated n-grams.

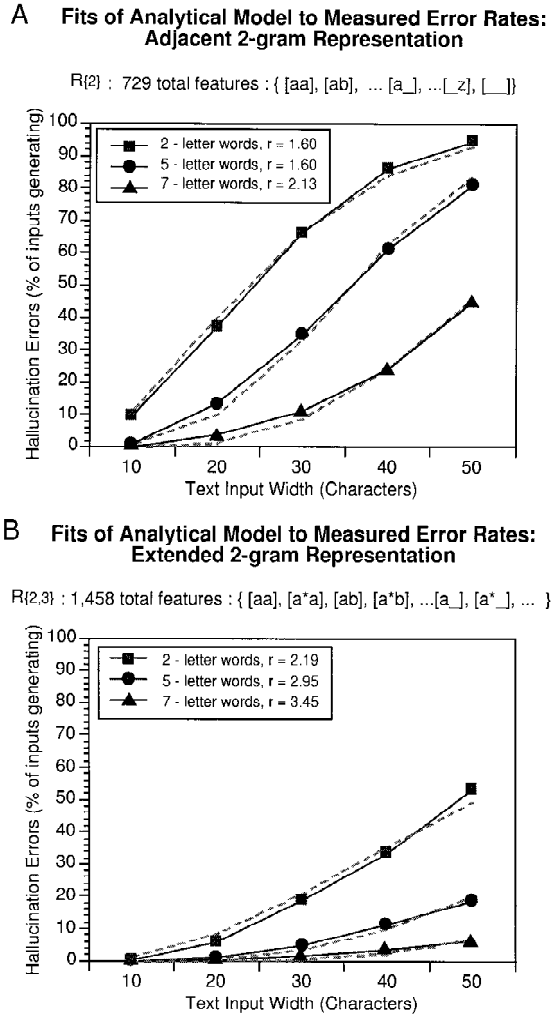


Figure 3: Fits of equation 4.3 to recognition performance in *text world*. (A) The representation, designated $\mathcal{R}_{\{2\}}$, contained the complete set of 729 adjacent 2-grams (including the space character). The x-axis shows the width of a text window presented to the n-gram representation, and the y-axis shows the probability that the input generates one or more hallucinated words. Analytical predictions are shown in dashed lines using redundancy factors specified in the legend. (B) Same as A, with results for $\mathcal{R}_{\{2,3\}}$.

2 and 7 letters, input arrays containing from 10 to 50 characters, and using representations varying in size by a factor of 2. We did not attempt to fit larger ranges of these experimental variables.

Regarding the sensitivity of the model fits to the choice of d' and r, we noted first that fits were not strongly affected by the cumulative probability cutoff used to choose d'—qualitatively reasonable fits could be also generated using a cutoff value of 95%, for example, for which d' values for $\mathcal{R}_{\{2,3\}}$ and $\mathcal{R}_{\{2\}}$ shrank to 220 and 504, respectively. In contrast, we noted that a change of one part per thousand in the value of r could result in a significant change in the shape a curve generated by equation 4.3, and with it significant degradation of the quality of a fit. However, the systematic increase in the optimal value of r with increasing word size and representation size indicates that this redundancy correction factor is not an entirely free parameter. It is also worth emphasizing that a redundancy "factor," which uniformly scales w, c, and d, is a very crude model of the complex statistical dependencies that exist among English n-grams; the fact that *any* set of small r values can be found that tracks empirical recognition curves over wide variations in the several other task variables is therefore significant.

5 Greedy Feature Learning

Several lessons from the analysis and experiments are that (1) larger representations can lead to better recognition performance, (2) only those features that are actually used should be included in a representation, (3) considerable statistical redundancy exists in a fully enumerated (or randomly drawn) set of conjunctive features, which should be suppressed if possible, and (4) features of low binding order, while potentially adequate to represent isolated objects, are insufficient for veridical perception of scenes containing multiple objects. In addition, we infer that different objects are likely to have different representational requirements depending on their complexity and that the composition of an ideal representation is likely to depend heavily on the particular recognition task. Taken together, these points suggest that a representation should be learned that includes higher-order features *only as needed*. In contrast, blind enumeration of an ever-increasing number of higher-order features is an expensive and relatively inefficient way to proceed.

Standard network approaches to supervised learning could in principle address all of the points raised above, including finding appropriate weights for features of various orders depending on their utility and helping to decorrelate the input representation. However, the potential need to acquire features of high order—so high that enumeration and testing of the full cohort of high-order features is impractical—compels the use of some form of explicit search as part of the learning process. The use of greedy, incremental, or heuristic techniques to grow a neural representation through increasing nonlinear orders is not new (e.g., Barron, Mucciardi, Cook, Craig,

& Barron, 1984; Fahlman & Lebiere, 1990), though such techniques have been relatively little discussed in the context of vision.

To cope with the many constraints that must be satisfied by a good visual representation, we developed a greedy supervised algorithm for feature learning that (1) selects the order in which features should be added to the representation, and (2) selects which features added to the representation should contribute to the minimal feature sets of each object in the database. By differentiating the log probability $u_v = \log(p_v) = \log \prod_i (1 - h_i)$ with respect to the hallucination probability of the ith object, we find that

$$\frac{du_v}{dh_i} = \frac{-1}{1 - h_i}, \tag{5.1}$$

indicating that the largest multiplicative increments to p_v are realized by squashing the largest values of $h_i = (c/d)^{w_i}$. Practically, this is achieved by incrementing the w-values (i.e., growing the minimal feature sets) of the most frequently hallucinated objects. The following learning algorithm aims to do this in an efficient way:

1. Start with a small bootstrap representation. (We typically used 50 2-grams found to be useful in pilot experiments.)
2. Present a large number of text "images" to the representation.
3. Collect hallucinated words—any word that was ever detected but not actually present—in an input image.
4. For each hallucinated word and the input array that generated it, increment a global frequency table for every n-gram (up to a maximum order of 5 and diameter of 6) that is (i) contained in the hallucinated word, (ii) not contained in the offending input, and (iii) not currently included in the representation.
5. Choose the most frequent n-gram in this table of any order and diameter, and add it to the current representation.
6. Build a connection from the newly added n-gram to any word detector involved in a successful vote for this feature in step 4. Inclusion of this n-gram in these words' minimal feature sets will eliminate the largest number of word hallucination events encountered in the set of training images.
7. Go to 2.

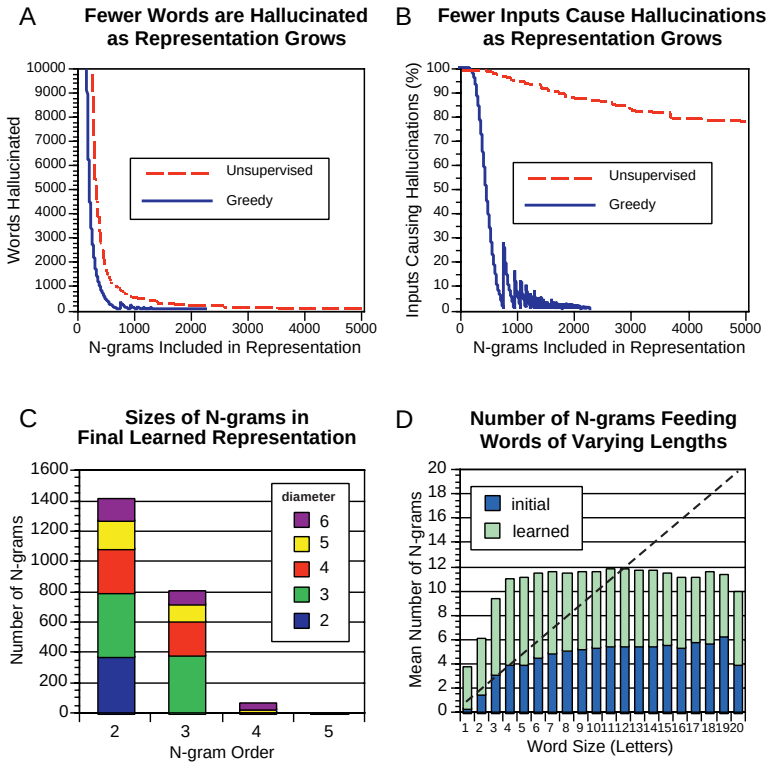
We ran the algorithm using input arrays 50 characters in width (often containing 10 or more words) and plotted the number of hallucinated words (see Figure 4A) and proportion of input arrays causing hallucinations (see Figure 4B) as a function of increasing representation size during learning. The periodic ratcheting in the performance curves (most visible in the lower curve of Figure 4B) was due to staging of the training process for efficiency

reasons: batches of 1000 input arrays were processed until a fixed hallucination error criterion of 1% was reached, when a new batch of 1000 random input arrays was drawn, and so on. The envelope of the oscillating performance curves provides an estimate of the true performance curve that would result from an infinitely large training corpus. Learning was terminated when the average initial error rate for three consecutive new training batches fell below 1%. Given that this termination condition was typically reached in fewer than 50 training epochs (consisting of 50,000 50-character input images), no more than half the text database was visited by random drawings of training-testing sets.

We found that for 50-character input arrays, the hallucination error rate fell below 1% after the inclusion of 2287 n-grams in $\mathcal{R}$. These results were typical. The inferior performance of a simple unsupervised algorithm, which adds n-grams to the representation in order of their raw frequencies in English, is shown in Figure 4AB for comparison (upper dashed curves). The counts of total words hallucinated (see Figure 4A), with initial values indicating many 10s of hallucinated words per input, can be seen to fall far faster than the hallucination error rate (see Figure 4B), since a sentence was scored as an error until its very last hallucinated word was quashed.

Figure 4: *Facing page*. Results of greedy n-gram learning algorithm using input arrays 50 characters in width and a 50-element bootstrap representation. Training was staged in batches of 1000 input arrays until the steady-state hallucination error rate fell below 1%. (A) Growth of the representation is plotted along the x-axis, and the number of words hallucinated at least once in the current 1000-input training batch is plotted on the y-axis. The drop is far faster for the greedy algorithm (lower curve) than for a frequency-based unsupervised algorithm (upper curve). Unsupervised bootstrap representation also included the 27 1-grams (i.e., individual letters plus space character) that were excluded from the bootstrap representation used in the greedy learning run. Without these 1-grams, performance of the unsupervised algorithm was even more severely hampered; *with* these 1-grams incorporated in the greedy learning runs, the w-values of every object were unduly inflated. (B) Same as (A), but the y-axis shows the proportion of input arrays producing at least one hallucination error. Jaggies occur at transitions to new training batches each time the 1% error criterion was reached. Asymptote is reached in this run after the inclusion of 2287 n-grams. (C) A breakdown of learned representation showing relative numbers of n-grams of each order and diameter. (D) The column height shows the minimal feature set size after learning, that is, average number of n-grams feeding word conjunctions for words of various lengths. The dark lower portion of the columns shows initial counts based on full connectivity to bootstrap representation; the light upper portions show an increment in w due to learning. Nearly constant w values after learning reflect the tendency of the algorithm to grow the smallest minimal feature sets. The dashed line corresponds to one n-gram per letter in word.

One of the most striking outcomes of the greedy learning process is that the learned representation is heavily weighted toward low-order features, as shown in Figure 4C. Thus, while the learned representation includes n-grams up to order 5, the number of higher-order features required to remedy perceptual errors falls off precipitously with increasing order—far faster than the explosive growth in the number of available features at each order. Quantitatively, the ratios of n-grams included in $\mathcal{R}$ to n-grams contained in the word database for orders 2, 3, 4, and 5 were 0.42, 0.0095, 0.00017, and 0.0000055, respectively. The rapidly diminishing ratios of useful features to available features at higher orders confirm the impracticality of learning on a fully enumerated high-order feature set and illustrate the natural bias in the learning procedure to choose the lowest-order features that "get the job done." The dominance of low-order features is of considerable practical significance, since higher-order features are more expensive to compute, less robust to image degradation, operate with lower duty cycles, and provide a more limited basis for generalization. The relatively broad distribution of



diameters of the learned feature set are also shown in Figure 4C for n-grams of each order.

In spite of its relatively small size, it is likely that the n-gram representation produced by the greedy algorithm could be significantly compacted without loss of recognition performance, by alternately pruning the least useful n-grams from $\mathcal{R}$ (which lead to minimal increments in error rate) and retraining back to criterion. However, such a scheme was not tried.

After learning to the 1% error criterion, a 50-character input array activated an average of 168 (7.3%) of the 2287 detectors in $\mathcal{R}$, while the average minimal feature set size for an individual word was 11.4. By comparison, the initial c/d ratio for 50-character inputs projected onto the 50-element bootstrap representation was a much less favorable 44%. Thus, a primary outcome of the learning algorithm in this case involving heavy clutter is a larger, more sparsely activated representation, which minimizes collisions between input images and the feature sets associated with individual target words. Further, the estimated r value from this run was 1.95, indicating significantly less redundancy in the learned representation than was estimated using seven-letter words ($r = 3.45$) in the fully enumerated 2-gram representations reported in Figure 3; this is in spite of the fact that the learned representation contained more than twice the number of active features as the fully enumerated 2-gram representation.

The total height of each column in Figure 4D shows the w value—the size of the minimal feature sets after learning for words of various lengths. The dark lower segment of each column counts initial connections to the word conjunctions from the 50-element bootstrap representation, while the average number of new connections gained during learning for words of each length is given by the height of each gray upper segment. Given the algorithm's tendency to expand the smallest minimal feature sets, the w values for short words grew proportionally much more than those for longer words, producing a final distribution of w values that was remarkably independent of word length for all but the shortest words. The w values for one-, two-, and three-letter words were lower either because they reached their maximum possible values (4 for one-letter words) or because these words contained rarer features, on average, derived from a relatively high concentration in the database of unpronounceable abbreviations, acronyms, and so forth. The dashed line indicates a minimal feature set size equal to the number of letters contained in the word; the longest words can be seen to depend on substantially fewer n-grams than they contain letters.

The pressures influencing both the choice of n-grams for inclusion in $\mathcal{R}$ and the assignment of n-grams to the minimal feature sets of individual words were very different from those associated with a frequency-based learning scheme. Thus, the features included in $\mathcal{R}$ were not nececssarily the most common ones, and the most commonly occurring English words were not necessarily the most heavily represented.

Table 3: First 10 n-Grams to Be Added to $\mathcal{R}$ During a Learning Run with 50-Character Input Arrays, Initialized with a 200-Element Bootstrap Representation.

Order of Inclusion	n-Gram	Relative Frequency	Cumulative %
1	[z_]	417	8
2	[_j]	14,167	55
3	[k*_]	39,590	71
4	[x_]	6090	42
5	[ki]	20,558	62
6	[_q]	8771	47
7	[_*v]	24,278	64
8	[_**z]	3184	31
9	[o*i]	37,562	70
10	[p***_]	55,567	76

Note: The "_" character represents the space character, and "*" matches any character. A value of k in the cumulative distribution column indicates that the total probability summed over all less common features is k%.)

First, in lieu of choosing the most commonly occurring n-grams, the greedy algorithm prefers features that are relatively rare in English text as a whole but relatively common in those words most likely to be hallucinated: short words and words containing relatively common English structure. The paradoxical need to find features that distinguish objects from common backgrounds, but do so as often as possible, leads to a representation containing features that occur with moderate frequency. Table 3 shows the first 10 n-grams added to a 200-element bootstrap representation during learning in one experimental run. The features' relative frequencies in English are shown, which jump about erratically, but are mostly from the middle ranges of the cumulative probability distribution for English n-grams. Most of these first 10 features include a space character, indicating that failure to represent elemental features properly in relation to object boundaries is a major contributor to hallucination errors in this domain. Specifically, we found that hallucination errors are frequently caused by the presence in the input of multiple "distractor" objects that together contain most or all of the features of a target object, for example, in the sense that *national* and *vacation* together contain nearly all the features of the word *nation*. To distinguish *nation* from the superposition of these two distractors requires the inclusion in the representation of a 2-gram such as [n*****_] whose diameter of 7 exceeds the maximum value of 6 arbitrarily established in the present experiments.

Having established that the features included in $\mathcal{R}$ are not keyed in a simple way to their relative frequencies in English, we also observed that no simple relation exists between word frequency and the per-word representational cost, that is, the minimal feature set size established during

Table 4: Correlation Between Word Frequency in English and Minimal Feature Set Size After Learning Was Only 0.1.

Word	Relative Frequency	Minimal Feature Set Size[a]
resting	100	25
leading	315	24
buffalo	188	8
unhappy	144	7
thereat	2	25
fording	1	23
bedknob	1	7
jackdaw	1	6

[a]Number of n-grams in $\mathcal{R}$ that feed the word's conjunction, that is, which must be present in order for a word to be detected. Table shows examples of common and uncommon words with large and small minimal features sets.

learning. In fact, if attention is restricted to the set of all words of a given length, with nominally equivalent representational requirements, the correlation between relative frequency of the word in English and its postlearning minimal feature set size was only 0.1. Thus some common words require a conjunction of many n-grams in $\mathcal{R}$, while others require few, and some rare words require many n-grams while others require few (see Table 4).

5.1 Sensitivity of the Learned Representation to Object Category Structure. Words are unlike common objects in that every word forms its own category, which must be distinguished from all other words. In *text world*, therefore, the critical issue of generalization to novel class exemplars cannot be directly studied. To address this shortcoming partially and assess the impact of broader category structures on the composition of the n-gram representation, we modified the learning algorithm so that a word was considered to be hallucinated only if no similar—rather than identical—word was present in the input when the target word was declared recognized. In this case, *similar* was defined as different up to any single letter replacement—for example, $\texttt{dog} = \{\texttt{fog}, \texttt{dig}, \ldots\}$, but $\texttt{dog} \neq \{\texttt{fig}, \texttt{og}, \texttt{dogs}, \ldots\}$. The resulting category structure imposed on words was thus a heavily overlapping one in which most words were similar to several other words.[4]

[4] This learning procedure, combined with the assumption that every word is represented by only a simple conjunction of features in $\mathcal{R}$ (disjunctions of multiple prototypes per word were not supported), cannot guarantee that every word detector is activated

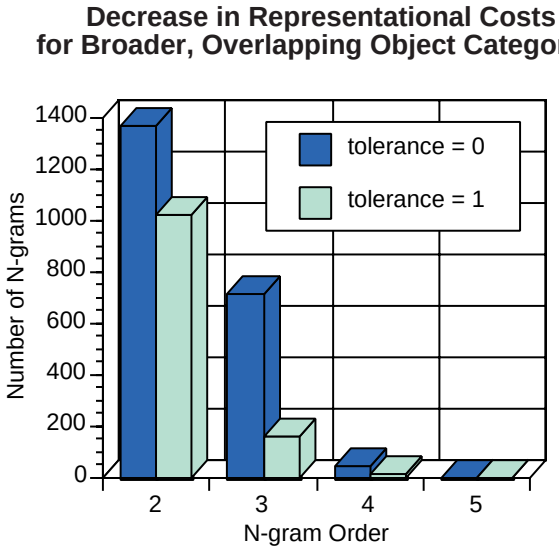


Figure 5: Breakdown of representations produced by two learning runs differing in breadth of category structure. Fifty-character input arrays were used as before to train to an asymptotic 1% error criterion, this time beginning with a 200-element bootstrap representation. Run with tolerance = 0 was as in Figure 4, where a hallucination was scored whenever a word's minimal feature set was activated, but the word was not identically present in the input. In a run with tolerance = 1, a hallucination was not scored if any similar word was present in the input, up to a single character replacement. The latter run, with a more lenient category structure, led to a representation that was significantly smaller, since it was not required to discriminate as frequently among highly similar objects.

The main result of using this more tolerant similarity metric during learning is that fewer word hallucinations are encountered per input image, leading to a final representation that is smaller by more than half (see Figure 5). In short, the demands on a conjunctive *n*-gram representation are lessened when the object category structure is broadened, since fewer distinctions among objects must be made.

5.2 Scaling of Learned Representation with Input Clutter. As is confirmed by the results of Figure 3, one of the most pernicious threats to a spatially invariant representation is that of clutter, defined here as the quantity of input material that must be simultaneously processed by the representa-

in response to *any* of the legal variants of the word. Rather, the procedure *allows* a word detector to be activated by any of the legal variants without penalty.

Table 5: Measured Average Values of c, d, and w and Calculated Values of r in a Series of Three Learning Runs with Input Clutter Load Increased from 25 to 75 Characters, Trained to a Fixed Error Criterion of 3%.

Input Width	c	d	w	r
25	55	816	7.6	1.45
50	152	1768	10.7	1.85
75	253	2634	12.8	2.11

Note: The 50-element bootstrap representation was used to initialize all three runs. Equation 4.3 was used to find values of r for each run, such that with $w \rightarrow w/r$, the measured values of c and d, and $N = 44{,}414$, a 3% error rate was predicted.

tion. The difficulty may be easily seen in equation 4.3, where an unopposed increase in the value of c leads to a precipitous drop in recognition performance. On the other hand, equation 4.3 also indicates that increases in c can be counteracted by appropriate increases in d and w, which conserve the value of $(c/d)^w$.

To examine this issue, we systematically increased the clutter load in a series of three learning runs from 25 to 75 characters, training always to a fixed 3% error criterion, and recorded for each run the postlearning values of c, d, and w. Although the WSJ database contained a wide range of word sizes (from 1 to 20 characters in length), the measurement of a single value for w averaged over all words was justified since, as shown in Figure 4D, the learning algorithm tended to equalize the minimal feature set sizes for words independent of their lengths. Under the assumption of uniformly probable, statistically independent features in the learned representations, the value of $h = (c/d)^w$ for all three runs should be equal to $\sim 7 \times 10^{-7}$, the value that predicts a 3% error rate for a 44,414-object database using equation 4.3. However, corrections ($w \rightarrow w/r$) were needed to factor out unequal amounts of statistical redundancy in representations of various sizes, where larger representations contained more redundancy.

The measured values of c, d, and w and the empirically determined r values are shown for each run in Table 5. As in the runs of Figure 3, redundancy values were again systematically greater for the larger representations generated, and the redundancy values calculated here for the learned representation were again significantly smaller than those seen for the fully enumerated 2-gram representation.

6 Discussion

Using a simple analytical model as a taking-off point and a variety of simulations in *text world*, we have shown that low-ambiguity perception in un-

segmented multiobject scenes can occur whenever the probability of fully activating any single object's minimal feature set, by a randomly drawn scene, is kept sufficiently low. One prescription for achieving this condition is straightforward: conjunctive features are mined from hallucinated objects until false-positive recognition errors are suppressed to criterion. Even in a complex domain containing a large number of highly similar object categories and severe background clutter, this prescription can produce a remarkably compact representation. This confirms that brute-force enumeration of large families or, worse, *all* high-order conjunctive features—the combinatorial explosion recognized by Malsburg—is not inevitable.

The analysis and experiments reported here have provided several insights into the properties of feature sets that support hallucination-proof recognition and into learning procedures capable of building these feature sets:

1. Efficient feature sets are likely to (1) contain features that span the range from very simple to very complex features, (2) contain relatively many simple features and relatively few complex features, (3) emphasize features that are only moderately common (giving a representation that is neither sparse nor dense) in response to the conflicting constraints that features should appear frequently in objects but not in backgrounds also composed of objects, and (4) in spatial domains, emphasize features that encode the relations of parts to object boundaries.

2. An efficient learning algorithm works to drive toward zero, and therefore to equalize, the false-positive recognition rates for all objects considered individually. Thus, frequently hallucinated objects—objects with few parts or common internal structure, or both—demand the most attention during learning. Two consequences of this focus of effort on frequently hallucinated objects are that (1) the average value of w, the size of the minimal feature set required to activate an object, becomes nearly independent of the number of parts contained in the object, so that simpler objects are relatively more intensively represented than complex objects, and (2) among objects of the same part complexity, the minimal conjunctive feature sets grow largest for objects containing the most common substructures, though these are not necessarily the most common objects. A curious implication of the pressure to represent objects heavily that are frequently hallucinated is that the backgrounds in which objects are typically embedded can strongly influence the composition of the optimal feature set for recognition.

3. The demands on a visual representation are heavily dependent on the object category structure imposed by the task. Where object classes are large and diffuse, the required representation is smaller and weighted

to features of lower conjunctive order, whereas for a category structure like words, in which every object forms its own category that must often be distinguished from a large number of highly similar objects (e.g., `cat` from `cats`, `cut`, `rat`), the representation must be larger and depend more heavily on features of higher conjunctive order.

6.1 Further Implications of the Analytical Model. Equation 4.3 provides an explicit mathematical relation among several quantities relating to multiple-object recognition. Since the equation assumes statistical independence and uniform activation probability of the features contained in the representation, neither of which is a valid assumption in most practical situations, we were unable to predict recognition errors using measured values for d, c, and w. However, we found that error rates in each case examined could be fitted using a small correction factor for statistical redundancy, which uniformly scaled d, c, and w and whose magnitude grew systematically for larger collections of features activated by larger—more redundant—units of text. From this we conclude that equation 4.3 at least crudely captures the quantitative trade-offs governing recognition performance in receptive-field-based visual systems.

One of the pithier features of equation 4.3 is that it provides a criterion for good recognition performance: $(c/d)^w \ll 1/N$. The exponential dependence on the minimal feature set size, w, means that modest increases in w can counter a large increase in the number of object categories, N, or unfavorable increases in the activation density, c/d, which could be due to either an increase in the clutter load or a collapse in the size of the representation. For example, using roughly the postlearning values of $c/d \sim 0.07$ and $r \sim 2$ from the run of Figure 4, we find that increasing w from 10 to 12 can compensate for a more than 10-fold increase N, or a nearly 60% increase in the c/d ratio, while maintaining the same recognition error rate.

The first-order expansion of equation 4.3 also states that recognition errors grow as the wth power of c/d, where w is generally much larger than 1. A visual system constructed to minimize hallucinations therefore abhors uncompensated increases in clutter or reduction in the size of the representation. These effects are exactly those that have motivated the various compensatory strategies discussed in section 1.

The abhorrence of clutter generates strong pressure to invoke any readily available segmentation strategies that limit processing to one or a small number of objects at a time. For example, cutting out just half the visual material in the input array when $w/r = 5$, as in the above example, knocks down the expected error rate by a factor of 32. The pressure to reduce the number of objects simultaneously presented to the visual system could account in part for the presence in biological visual systems of (1) a fovea, which leads to a huge overrepresentation of the center of fixation and marginalization of the surround, (2) covert attentional mechanisms, which selectively up- or down-modulate sensitivity to different portions of the visual field, and

(3) dynamical processes that segment "good" figures (in the Gestalt sense) from background.

The second pressure involved in maintaining a low c/d ratio involves maintaining a visual representation of adequate size, as measured by d. One situation in which this is particularly difficult arises when the task mandates that objects be distinguished under excessively wide ranges of spatial variation, an invariance load that must be borne in turn by each of the individual feature detectors in the representation. For example, consider a relatively simple task in which the orientation of image features is a valid cue to object identity, such as a task in which objects usually appear in a standard orientation. In this case a bank of several orientation-specific variants of each feature can be separately maintained within the representation. In contrast, in a more difficult task that requires that objects be recognized in any orientation, each bank of orientation-specific detectors must be pooled together to create a single, orientation-invariant detector for that conjunctive feature. The inclusion of this invariance in the task definition can thus lead to an order-of-magnitude reduction in the size of the representation, which, unopposed could produce a catastrophic breakdown of recognition according to equation 4.3. As a rule of thumb, therefore, the inclusion of an additional order-of-magnitude invariance must be countered by an order-of-magnitude increase in the number of distinct feature conjunctions included in the representation, drawn from the reservoir of higher-order features contained in the objects in question. The main cost in this approach lies in the additional hardware, requiring components that are more numerous, expensive, prone to failure, and used at a far lower rate.

6.2 Space, Shape, Invariance, and the Binding Problem. While one emphasis in this work has been the issue of spatial invariance in a visual representation, we reiterate that the mathematical relation expressed by equation 4.3 knows nothing of space or shape or invariance, but rather views objects and scenes as collections of statistically independent features. What, then, is the role of space?

From the point of view of predicting recognition performance and with regard to the quantitative trade-offs discussed here, we can find no special status for worlds in which spatial relations among parts help to distinguish objects, since spatial relations may be viewed simply as additional sources of information regarding object identity. (Not all worlds have this property; an example is olfactory identification.) Further, the issue of spatial invariance plays into the mathematical relation of equation 4.3 only indirectly, in that it specifies the degree of spatial pooling needed to construct the invariant features included in the representation; once again, the details relating to the internal construction of these features are not visible to the analysis. In this sense, the spatial binding problem, which appears at first to derive from an underspecification of the spatial relations needed to glue an object together, in fact reduces to the more mundane problem of ambiguity arising

from overstimulation of the visual representation by input images—that is, a c/d ratio that is simply too large. The hallucination of objects based on parts contained in background clutter can thus occur in any world, even one in which there is no notion of space (e.g., matching bowls of alphabet soup).

However, the fact that object recognition in real-world situations typically depends on spatial relations between parts and the fact that recognition invariances are also very often spatial in nature together strongly influence the way in which a visual representation based on spatially invariant, receptive fields can be most efficiently constructed. In particular, many economies of design can be realized through the use of hierarchy, as exemplified by the seminal architecture of Fukushima et al. (1983). The key observation is that a population of spatially invariant, higher-order conjunctive features generally shares many underlying processing operations. The present analysis is, however, mute on this issue.

6.3 Relations to "Real" Vision. The quantitative relations given by equation 4.3 cast the problem of ambiguous perception in cluttered scenes in the simple terms of set intersection probabilities: a cluttered scene activates c of the d feature detectors at random and is correctly perceived with high probability as long as the set of c activated features only rarely includes all w features associated with any one of the N known objects. The analysis is thus heavily abstracted from the details of the visual recognition problem, most obviously ignoring the particular selectivites and invariances that parameterize the detectors contained in the representation. On the other hand, the analysis makes it possible to understand in a straightforward way how the simultaneous challenges of clutter and invariance conspire to create binding ambiguity and the degree to which compensatory increases in the size of the representation, through conjunctive order boosting, can be used to suppress this ambiguity. These basic trade-offs have operated close to their predicted levels in the *text world* experiments carried out in this work, in spite of violations of the simplifying assumptions of feature independence and equiprobability.

Since the application of our analysis to any particular recognition problem involves estimating the domain-specific quantities c, d, w, and N, we do not expect predictions relating to levels of binding ambiguity or recognition performance to carry over directly from the problem of word recognition in blocks of symbolic text to other visual recognition domains, such as viewpoint-independent recognition of real three-dimensional objects embedded in cluttered scenes.

On the other hand, we expect the basic representational trade-offs to persist. The most important way that recognition in more realistic visual domains differs from recognition in *text world* lies in the invariance load, which in most interesting cases extends well beyond simple translation invariance. Basic-level object naming in humans, for example, is largely invariant to

translation, rotation, scale, and various forms of distortion, occlusion, and degradation—invariances that persist even for certain classes of nonsense objects (Biederman, 1995). Following the logic of our probabilistic analysis, the daunting array of invariances that must be maintained in this and many other natural visual tasks would seem to present a huge challenge to biological visual systems. In keeping with this, performance limitations in human perception suggest that our visual systems have responded in predictable ways to the pressures of the various tasks with which they are confronted; in particular, recognition in a variety of visual domains appears to include only those invariances that are absolutely necessary and those for which simple compensatory strategies are not available. For example (1) recognition is restricted to a highly localized foveal acceptance window for text or other detailed figures, where by "giving up" on translation invariance, this substream of the visual system frees hardware resources for the several other essential invariances that cannot be so easily compensated for (e.g. size, font, kerning, orientation, distortion), (2) face recognition operates under strong restrictions on the image-plane orientation and direction of lighting of faces (Yin, 1969; Johnston, Hill, & Carman, 1992), a reasonable compromise given that faces almost always present themselves to our visual systems upright and with positive contrast and lighting from above, and (3) unlike the case for common objects, reliable discrimination of complex three-dimensional nonsense objects (e.g., bent paper clips, crumpled pieces of paper) is restricted to a very limited range of three-dimensional orientations in the vicinity of familiar object views (Biederman & Gerhardstein, 1995), though this deficiency can be overcome in some cases by extensive training (Logothetis & Pauls, 1995). When viewed within our framework, the exceptional difficulty of this latter task arises from the need for, or more precisely a lack of, the very large number of very-high-order conjunctive features—essentially full object views locked to specific orientations (see Figure 1B)—that are necessary to support reliable viewpoint-invariant discrimination among a large set of objects with highly overlapping part structures. In summary, the primate visual system manifests a variety of domain-specific design compromises and performance limitations, which can be interpreted as attempts to achieve simultaneously the needed perceptual invariances, acceptably low levels of binding ambiguity, and reasonable hardware costs.

In considering the astounding overall performance of the human visual system in comparison to the technical state of the art, it is also worth noting that human recognition performance is based on neural representations that could contain tens of thousands of task-appropriate visual features (extrapolated from the monkey; see Tanaka, 1996) spanning a wide range of conjunctive orders and finely tuned invariances.

In continuing work, we are exploring further implications of the tradeoffs discussed here in relation to the neurobiological underpinnings of spatially invariant conjunctive receptive fields in visual cortex (Mel, Ruderman & Archie, 1998) and on the development of more efficient supervised

and unsupervised hierarchical learning procedures needed to build high-performance, task-specific visual recognition machines.

Acknowledgments

Thanks to Gary Holt for many helpful comments on this work. This work was funded by the Office of Naval Research.

References

Barron, R., Mucciardi, A., Cook, F., Craig, J., & Barron, A. (1984). Adaptive learning networks: Development and Applications in the United States of algorithms related to GMDH. In S. Farrow (Ed.), *Self-organizing methods in modeling*. New York: Marcel Dekker.

Biederman, I. (1995). Visual object recognition. In S. Kosslyn & D. Osherson (Eds.), *An invitation to cognitive science* (2nd ed.) (pp. 121–165). Cambridge, MA: MIT Press.

Biederman, I., & Gerhardstein, P. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995). *J. Exp. Psychol. (Human Perception and Performance), 21*, 1506–1514.

Califano, A., & Mohan, R. (1994). Multidimensional indexing for recognizing visual shapes. *IEEE Trans. on PAMI, 16*, 373–392.

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.

Douglas, R., & Martin, K. (1998). Neocortex. In G. Shepherd (Ed.), *The synaptic organization of the brain* (pp. 567–509). Oxford: Oxford University Press.

Edelman, S., & Duvdevani-Bar, S. (1997). A model of visual recognition and categorization. *Phil. Trans. R Soc. Lond. [Biol], 352*, 1191–1202.

Fahlman, S., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. Touretzky (Ed.), *Advances in neural information processing systems, 2* (pp. 524–532). San Mateo, CA: Morgan Kaufmann.

Fize, D., Boulanouar, K., Ranjeva, J., Fabre-Thorpe, M., & Thorpe, S. (1998). Brain activity during rapid scene categorization—a study using event-related FMRI. *J. Cog. Neurosci., suppl S*, 72–72.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognition: A neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Sys. Man & Cybernetics, SMC-13*, 826–834.

Gilbert, C. (1983). Microcircuitry of the visual cortex. *Ann. Rev. Neurosci, 89*, 8366–8370.

Heller, J., Hertz, J., Kjær, T., & Richmond, B. (1995). Information flow and temporal coding in primate pattern vision. *J. Comput. Neurosci., 2*, 175–193.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol., 195*, 215–243.

Hummel, J., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psych. Rev., 99*, 480–517.

Johnston, A., Hill, H., & Carman, N. (1992). Recognizing faces: Effects of lighting direction, inversion, and brightness reversal. *Perception, 21*, 365–375.

Jones, E. (1981). Anatomy of cerebral cortex: Columnar input-output relations. In F. Schmitt, F. Worden, G. Adelman, & S. Dennis (Eds.), *The organization of cerebral cortex*. Cambridge, MA: MIT Press.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., Malsburg, C., Wurtz, R., & Komen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers, 42*, 300–311.

Lang, G. K., & Seitz, P. (1997). Robust classification of arbitrary object classes based on hierarchical spatial feature-matching. *Machine Vision and Applications, 10*, 123–135.

Le Cun, Y., Matan, O., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., & Baird, H. (1990). Handwritten zip code recognition with multilayer networks. In *Proc. of the 10th Int. Conf. on Patt. Rec*. Los Alamitos, CA: IEEE Computer Science Press.

Logothetis, N., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex, 3*, 270–288.

Logothetis, N., & Sheinberg, D. (1996). Visual object recognition. *Ann. Rev. Neurosci., 19*, 577–621.

Malsburg, C. (1994). The correlation theory of brain function (reprint from 1981). In E. Domany, J. van Hemmen, & K. Schulten (Eds.), *Models of neural networks II* (pp. 95–119). Berlin: Springer.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception. *Psych. Rev., 88*, 375–407.

Mel, B. W. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation, 9*, 777–804.

Mel, B. W., Ruderman, D. L., & Archie, K. A. (1998). Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. *J. Neurosci., 17*, 4325–4334.

Mozer, M. (1991). *The perception of multiple objects*. Cambridge, MA: MIT Press.

Oram, M., & Perrett, D. (1992). Time course of neural responses discriminating different views of the face and head. *J. Neurophysiol., 68*(1), 70–84.

Oram, M. W., & Perrett, D. I. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks, 7*(6/7), 945–972.

Pitts, W., & McCullough, W. (1947). How we know universals: The perception of auditory and visual forms. *Bull. Math. Biophys., 9*, 127–147.

Potter, M. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol.: Human Learning and Memory, 2*, 509–522.

Sandon, P., & Urh, L. (1988). An adaptive model for viewpoint-invariant object recognition. In *Proc. of the 10th Ann. Conf. of the Cog. Sci. Soc.* (pp. 209–215). Hillsdale, NJ: Erlbaum.

Schiele, B., & Crowley, J. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *Proc. of the 13th Int. Conf. on Patt. Rec.* (Vol. 2 pp. 50–54). Los Alamitos, CA: IEEE Computer Society Press.

Swain, M., & Ballard, D. (1991). Color indexing. *Int. J. Computer Vision, 7*, 11–32.

Szentagothai, J. (1977). The neuron network of the cerebral cortex: A functional interpretation. *Proc. R. Soc. Lond. B, 201*, 219–248.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Ann. Rev. Neurosci, 19*, 109–139.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Van Essen, D. (1985). Functional organization of primate visual cortex. In A. Peters & E. Jones (Eds.), *Cerebral cortex* (pp. 259–329). New York: Plenum Publishing.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol., 51*, 167–194.

Weng, J., Ahuja, N., & Huang, T. S. (1997). Learning recognition and segmentation using the cresceptron. *Int. J. Comp. Vis., 25*(2), 109–143.

Wickelgren, W. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psych. Rev., 76*, 1–15.

Yin, R. (1969). Looking at upside down faces. *J. Exp. Psychol., 81*, 141–145.

Zemel, R., Mozer, M., & Hinton, G. (1990). TRAFFIC: Recognizing objects using hierarchical reference frame transformations. In D. Touretzky (Ed.), *Advances in neural information processing systems, 2* (pp. 266–273). San Mateo, CA: Morgan Kaufmann.

Received September 29, 1998; accepted March 5, 1999.