Crowding: A neuroanalytic approach

Christopher W. Tyler

The Smith-Kettlewell Eye Research Institute, San Francisco, CA, USA

Lora T. Likova

The Smith-Kettlewell Eye Research Institute, San Francisco, CA, USA

The mechanisms underlying crowding are analyzed in terms of explicit neural processes mediating its perceptual characteristics as originally described by W. Korte (1923). A full understanding of crowding in letter recognition requires a detailed conceptualization of the process of recognition among large numbers of alternatives. The observed masking properties suggest the operation of recursive inhibition from V3 to V1 as a component of the crowding effect. The plausibility of six accounts of the neural basis of crowding (the template matching, feature integrator, attentional feature conjunction, propositional enumeration, attentional tracking, and relaxation network concepts) is then assessed in relation to the task of encoding the spatial structure of the letter forms. We conclude that the relaxation network approach is the most plausible hypothesis to account for the full-spectrum letter recognition performance.

Keywords: crowding, visual acuity, attention, spatial integration, letter recognition, reading, perceptual organization

Citation: Tyler, C. W., & Likova, L. T. (2007). Crowding: A neuroanalytic approach. *Journal of Vision, 7*(2):16, 1–9, http://journalofvision.org/7/2/16/, doi:10.1167/7.2.16.

Introduction

Crowding is defined as impaired recognition of a suprathreshold target due to the presence of distractor elements in the neighborhood of that target. The classic example of crowding is the case of letter discrimination rendered impossible by the presence of neighboring letters. The editors (Pelli, Cavanagh, Desimone, Tjan, & Treisman, 2007) invited consideration of the competing analyses in terms of a spatial processing of cluttered information (Pelli, Palomares, & Majaj, 2004) versus a local limitation on the attentional readout of information that is neurally resolved at the lower levels of processing (Intriligator & Cavanagh, 2001). To do so, the prime requirement is to understand what kind of process is involved in the operation of the feature integration that underlies the letter recognition task. Examination of such analyses of the properties of crowding reveals that they do not attempt to uncover the neural processes required to perform the recognition task. Both approaches assume a comparator of some kind that generates the decision as to which letter was present, and measure the properties of its range of operation without addressing the mechanism by which the recognition itself takes place.

The goal of this article is to explore the neural processes required to account for the crowding phenomenon and to unmask the implicit homuncular assumption embedded in the concept of a feature integrator or attentional comparator that distinguishes among visual items made up of similar elements. To do so, we need to consider how the crowded stimulus actually looks to the viewer and the implications of this phenomenology for the competing explanations of the crowding effect. We then review several accounts of process of letter recognition and provide a neuroanalytically plausible account of the mechanisms by which the visual system can perform the complex discriminations involved in letter recognition.

Demonstration

Figure 1 shows an example of a crowded stimulus. To provide your own phenomenology of its discriminability, fixate steadily on the plus sign and assess the visibility of the letters below (avoiding the temptation to move the eyes).

The typical report is that, while fixating is maintained on the cross, the two isolated letters at the left and right are easily read as X and N and the two flanking letters below the vertical line are relatively visible, but the center of the three symbols is heavily obscured by its adjacent letters. Notice that it is even hard to be sure that there is a letter there at all, or whether it has an oblique element in it.

A common description of the phenomenology of crowding is exemplified by:

> Each letter may seem at times to be an A and sometimes a B, but most of the time it has a confusing hybrid A–B appearance that would be impossible to draw (Pelli et al., 2004).

doi: 10.1167/7.2.16 Received March 15, 2007; published July 30, 2007 ISSN 1534-7362

Such accounts emphasize the *confusion* aspect of crowding, implying that the local feature elements are fully visible, but their relative locations are confused, making it impossible to identify the letter in the center of the triplet (as depicted in Figure 2a). Careful observation during fixation of Figure 1, however, reveals that crowded the elements themselves are very unclear. The predominant impression is that there is a gray, or inchoate, smudge between the two outer letters, including the inner parts of those letters. The same point was made by Korte (1923), the original discoverer of the crowding phenomenon:

> It is as if there is a pressure on both sides of the word that tends to compress it. Then the stronger, i.e. the more salient or dominant letters, are preserved and they "squash" the weaker, i.e., the less salient letters, between them (Korte, 1923, translated by Uta Wolfe).

Thus, the crowded visual impression is closer to the depiction of Figure 2b than that of Figure 2a. This result suggests that there is some inhibitory process in operation, beyond the aspect of spatial uncertainty implied by the "scrambling" description.

Analysis

We can use Korte's (1923) specification of crowding to analyze the neural mechanisms involved. Consider first the perceived masking in crowding, as depicted in Figure 2b. This reduction in visibility of the elementary features can be understood in terms of the known connections between cells in the sequence of retinotopic maps through which visual information passes after arriving at V1. It is well established that the receptive field sizes increase by successive factors of about 2 from V1 > V2 > V3 > V3A/V4 (Zeki, 1978). If the letters are set to match the scale of the V1 receptive fields, the flanking letters at a distance of only one letter spacing should have no effect in V1 in terms of the classical receptive fields. They will, however, fall within the receptive fields of V2 and V3 keyed to the location of the test letter. The other property of the retinotopic maps is that they exhibit extensive recurrent inhibition between maps. Thus, the activation of the V2/V3 maps by the flanking letters will result in inhibitory

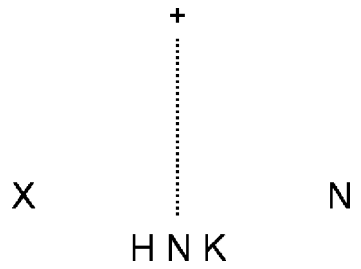

Figure 1. A demonstration of the crowding effect.

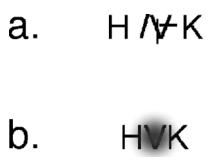

Figure 2. (a) Depiction of the uncertainty description of the crowding percept. (b) Depiction of Korte's (1923) original description of the crowding percept.

suppression of the neurons responding to the target letters in V1. The suppression is not likely to be complete but to induce sufficient inhibitory reduction to account for the reduced visibility noted by Korte. The scale of the crowding effect, which is given by Pelli et al. (2004) as approximately 0.4 E at any peripheral eccentricity E, is such as to require the inhibitory fields to be located in V3, based on the quantification of its size scale. We take the operation of the inhibition to be a form of "contrast contrast", the reduction in perceived contrast and elevation of detection threshold of a region when surrounded by high-contrast texture relative to their perception with a uniform surround (Cannon & Fullenkamp, 1991; 1993; Chubb, Sperling, & Solomon, 1989; Ejima & Takahashi, 1985; Ellemberg, Wilkinson, Wilson, & Arsenault, 1998; McDonald & Tadmor, 2006; Snowden & Hammett, 1998; Xing & Heeger, 2000, 2001). Such results can be explained by a quantitative spatial normalization model of the dual masking and sensitivity modulation of the visibility of a central target by flanking elements (Chen & Tyler, 2001, 2002). It was mentioned in the Introduction section that previous analyses of the mechanism of crowding (Intriligator & Cavanagh, 2001; Pelli et al., 2004; Põder, 2006) do not provide a complete account of the recognition process. Specifically:

> Attention appears to speed processing of objects that are part of the same surface regardless of their absolute spatial location… Given the range of views concerning the mechanisms of attention, it is hard to describe the region over which attention operates in terms compatible with all the alternatives—space- or object-based selection, resource allocation, or filters. For simplicity, we use the term "selection" to describe the operation of attention and "region of selection" to describe the area over which it operates but we do not favor one model over another (Intriligator & Cavanagh, 2001, p. 173).

> Despite progress in vision research, we still can only barely begin to answer a simple question like, "How do I recognize the letter A?" …This (nonlinear) assembly process is called "feature integration" (or "binding"). Feature integration may internally represent the combined features as an object, but we will not address that here (Pelli et al., 2004, p. 1137).

Thus, both papers accept the idea that the signals for local features must be combined to produce the decision

as to which letter is present at the test location but skirt the issue of the neural mechanism by which the combination of features and their relative locations takes place.

To understand a perceptual process, the position of the present neuroanalytic treatment is that one must be able to specify the processing circuitry in terms of known functions of cortical neurons, generating an output from a neural array that performs the task in a manner that can be implemented in a computational simulation. Explanations expressed in terms such as "attentional integration fields", "decisions", and "confusions" are nonspecific processes expressed in the language of the mind. For a neuroanalytic explanation, on the other hand, the relevant concepts are nonlinear receptive fields, neural spikes, and noisy signals, respectively.

The non-neurally specific reproach even applies to the concept of a "feature integrator" (Itti & Koch, 2001; Pelli et al., 2004; Põder, 2006; Treisman & Gelade, 1980) or, equivalently, an "attentional selection mechanism" (Intriligator & Cavanagh, 2001). The role of these attentional integrators or selectors is to determine which of many alternative combinations of symbols was presented at a given (crowded) location in the field.

> A pointer or index could then indicate the location of the selected items, say, in area V1. When required for further analysis, the properties of the item at that location could be read out via the pointer. In this case, there is no need for a higher cortical area to represent the full range of orientations, sizes, and colors that we might experience from the attended object (Intriligator & Cavanagh, 2001, p. 210).

This concept of a set of pointers in parietal cortex indexing objects or regions of interest in early visual areas is similar to Pylyshyn's (1989, 2001) "fingers of instantiation", whose function was to highlight objects of interest.

However, these concepts are not offered as a set of hardwired neural receptive fields with a mechanism for determining the one with the maximal response or other known neural implementations. They are presented as quasi-intelligent analyzers, able to assess the conjunction of any pair (or combination) of features. Thus, the essence of the definition of integration fields in Pelli et al. (2004) is contained in the following quote:

> if a … complex judgment is required, the observer [*sic*] may monitor the output of a feature integrator whose integration field has a minimum size defined by the critical spacing, allowing a response based on a combination of the features in that integration field. … It might seem fanciful to make strong assertions about "integration fields" that are so vaguely specified…. Receptive fields detect features. Such feature-detection events are subsequently combined by integration fields (Pelli et al., 2004, pp. 1155–1156).[1]

This quote slips between the terms "feature integrator" and "integration field", but let us focus on the intent, which is to discuss the mechanism recognizing the feature combination, which must be the "feature integrator". Invocation of a metaphorical "observer" is an explicitly homuncular form of the neural monitoring mechanism that makes a decision based on the feature linkage process. The authors of the quote appear to recognize the non-analytic nature of their approach in characterizing it as 'fanciful' and 'vaguely specified'. If we translate the quote into neural language, we understand that the authors are delineating a neural mechanism that receives the outputs from a (large?) set of neurons with local receptive fields selective for all the possible features that the experimenters may have decided to present and then to "combine" them in some (unspecified) fashion that delivers a meaningful discriminative signal to guide the actual observer's behavior. This description is typical of those who invoke the concepts of integration or attentional fields in recognition tasks. They imply that it is possible for the feature integrator to perform the requisite combination, as though it had hands and could link the input signals in some lego-like fashion, without specifying the neural processing by which such selective linkage could occur. With this caveat in mind, we attempt to specify the details of the letter discrimination task performed by humans at a level that could permit the analysis of the mechanisms involved in strictly neural terms.

Letter discrimination

The typical task used to examine crowding is a multiple letter discrimination with letters that are suprathreshold and readily discriminable in isolation. Letters consist of about six simple elements combined in a variety of lengths and arrangements (horizontal, vertical, left oblique, right oblique, small curve, and large curve), with repeats allowed. This set makes up a complex vocabulary of shapes, with 26 options in the English language, but on the order of 2^8 easily recognizable icons on the computer keyboard (when other symbols and alphabets such as Greek and Cyrillic are included).

The first question, therefore, is how do we perform such a complex task? What neural architecture permits the discrimination among such a large number of alternatives? Until their acuity limit is reached, it is easy for most people to perform letter recognition at an accuracy rate of 99% or better (e.g., when reading). Many theories of letter discrimination simply assume that, if all the *elements* of the letter are encodable by local receptive fields, the task of discriminating among their *combinations* is perceptually solved. This homuncular assumption, that recognition is straightforward once the elements are discriminable, begs the question of how the combinations are encoded as such. The logic of neural analysis implies that there must be some circuit somewhere in the brain that embodies the

concept of each letter or symbol *exclusive of all other symbols,* together with its name and routines for its verbal pronunciation (in most cases). For each recognition event, this particular circuit must be activated above all other circuits encoding the other recognized items.

Encoding, labels, conjunctions, and attention

There are basically three proposals as to how letter encoding could take place. One is that there could be *templates* for each letter type (or combination of elementary features), such that the best fitting template generates the largest response for the set and acts as a "labeled line" for the identified letter type. The template model is usually rejected on the grounds that it requires an implausibly large number of templates to match all the letters in all the combinations of fonts, sizes, and rotations required to account for the typical adult letter recognition performance. Nevertheless, it has been successfully implemented for handwriting recognition (Larsen & Bundesen, 1996), although its performance under crowded conditions was not addressed.

The concepts of an "integration field" or "attentional selection" constitute a second approach, intended to circumvent the difficulties of template matching by positing the existence of a neural operator sensitive to the combinations of elements and able to signal its choice of solution to the combination problem represented by the letter discrimination task. As far as we are aware, the neural process by which this choice takes place has not been addressed by any of its proponents. Of course, we know that the brain as a whole can perform such categorical operations, but that is equivalent to the homuncular assumption that the neural feature detectors deliver their labeled signals to consciousness, where the large-scale cell assemblies "know" how to perform the categorical operations required for the discrimination. Is this what the proponents mean by the "feature integrator"?

We suggest that a much more helpful term would be "comparator mechanism", which at least specifies the function required of this second-level neural process. The concept of "integration" seems far off the mark because integration is well established as a term for summation (linear or nonlinear), but summation simply incorporates the signals from two or more synaptic inputs with no room for labels as to their individual identities; thus, it simply generates stronger or more selective activation for the conjunction than for the individual elements. Unless there is a separate "feature integrator" for each combination, the integration concept has advanced no further in categorizing the stimulus. None of the papers that we cite have proposed such a discrete set of comparators. If they did so, they would effectively be proposing a set of (possibly nonlinear) templates, one for each possible feature combination, and be subject again to the implicit objections against template matching.

In practice, many studies are designed to pare down the choice of letters to just two, such as L and T or X and O, which could be performed by a simple discriminator that does not place such exorbitant demands on the comparator mechanism. This simplification may be the reason why the investigators in the field do not feel the need to specify how the "feature integrator" performs its categorization operation. However, crowding is a phenomenon that potentially operates in every eye clinic every time someone reads an eye chart. Unless the letters are sufficiently well spaced, they are subject to the crowding constraint that would distort the results in relation to the theoretical acuity measurement delivered by isolated letters. Such crowding uses a large vocabulary of letters or symbols, and this is the complex task that ultimately needs to be explained. The third approach, deriving from Treisman and Gelade (1980), is that the feature conjunctions required for recognition are performed by "attention". In their model, the conjunction locates the two features in separate "feature maps", reporting a conjunction "hit" if the locations match (and an "illusory conjunction" if the match was spurious). This theory does not account for the ability to find two or more such conjunctions in a visual search task. To do so, it seems that there must be a wide-field attentional monitoring mechanism that can focus on the most salient feature in the field of view at any moment (Kontsevich & Tyler, 1999; Lee, Itti, Koch, & Braun, 1999; Parkhurst, Law, & Niebur, 2002).

The attention concept was extended to spatial conjunctions by Intriligator and Cavanagh (2001). They are not explicit about whether the conjunctions are identified serially or in parallel, but it is hard to conceive that overlearned tasks such as letter recognition would be considered to require serial comparisons, especially when letter recognition operates at the pace of ~10 ms per item (Sperling, Budiansky, Spivak, & Johnson, 1971). Thus, we assume that the application of attention models in the crowding paradigm is intended as a parallel comparator.

But how do the proponents conceive "attention" to operate as a comparator? The core concept of attention is to enhance selectivity of the outputs of particular neural channels. The selection may be driven by stimulus relationships (exogenous attention) or by previous neurally encoded preferences (endogenous attention). In either case, selection per se does not implement a comparator role. By assigning the comparator role to "attention", its proponents are expanding the capabilities of that enhancement process to include conjunction, localization, memory access, and recognition. In each case, they are begging the question of how these process components are instantiated neurally. For example, a Treisman-like feature conjunction cannot discriminate between an "L" and a "T" because both have the same combination elementary of features. Although the measurements reveal that there is a selection process for targets in the visual field, and that selection process has a limited spatial resolution, the neural mechanism for the

selection operation and the reason for its resolution limit remain obscure. If "attention" can select items that are an order of magnitude smaller than its field of selection, what is it that limits the selection access *within* the attention field?

One can invoke a physical analogy, that one can pick up items much smaller than one's fingers (e.g., a grain of rice), but one cannot select a short grain from a dense field of long grains with any degree of accuracy. Thus, the selected item can be much finer than the ability to select among items. However, if this kind of attention limit is part of a theory of object processing, it has to be explained by what mechanism the properties of the selected grain (long or short grain?) are assessed. In the case of the fingers, it is the array of touch receptors in the skin, allowing detailed information about the grain size and shape to be encoded at a scale much smaller than the fingertip. This analogy helps to clarify that a low-resolution attention mechanism still requires complex internal structure to be able to perform the conjunction task on the array of potential features making up each discriminandum.

Now, the metaphor of the fingertip being used to estimate the shape of the grain must be translated into a neural mechanism for performing the discrimination in the letter recognition task. To do so for an acuity chart, there must be some kind of coding of the orientation and position of each line of letter symbols. This concatenation of features must then be compared with sets of memory traces to generate the recognition response (which we assume to be verbal, as in the standard use of an acuity chart). According to Treisman (1998, pp. 196–197):

> The attended features can then be entered, without risk of binding errors, into the currently active object representation where their structural relations can be analysed.

Thus, Treisman's concept focuses on the selection and conjunction per se but has no explicit neural mechanism for analyzing the relationships among the conjoined elements beyond the specification that it takes place in the "currently active object representation".

In summary, the template matching, feature integrator, and attentional feature conjunction hypotheses all fall short of specifying the dynamic comparator operations implied by human performance in reading the common eye chart. Having established that there must be a dynamic comparator operation underlying letter recognition performance, we now consider three models of how this comparison may take place: *propositional enumeration* of the logical information in the visual field, *attentional tracking* of the paths defined the visual elements, and *network relaxation minima* encoding the spatial structure of the image.

Discussion

Propositional enumeration

The propositional enumeration model corresponds to Pylyshyn's (1999) concept of perceptual processing, which is that we "make sense" of sensory information by converting it to a propositional code that is closely linked to the verbal description of the objects in the scene and the relations to ourselves and each other. In terms of letter recognition, the role of attention would be to select the particular items in the scene to be highlighted for propositional coding at any given moment. This view emphasizes not only the local spotlight of attention but also the beam of information flowing back from the highlighted region, allowing recognition processing by accessing comparative structures stored in long-term memory. In Pylyshyn's view, the mode of comparison for feature conjunctions is *propositional,* in the sense of operating with relations among encoded lists of properties rather than image-space templates. Thus, a "Y" would be encoded as having the properties "vertical below, left-oblique above left, right-oblique above right" with respect to the letter center. This (preverbal) code would then be used to search for matching combinations in the memory bank, presumably by some prioritized lexical access scheme analogous to the organization of a dictionary. The operation of the Google search engine with a similar strategy makes it plausible that such a scheme could function in the real time available for human letter recognition.

How would the propositional recognition scheme account for the phenomena of crowding? The implication is that the features of adjacent letters would throw in extra propositional items within the attentional spotlight that would disrupt the recognition process. What this explanation assumes is that, because there is no reason for the feature identities to be disrupted by flanking letters, the disruption must occur in the *place coding* for each elementary feature (e.g., "above left"). If there are too many items within the attentional pipeline, there is insufficient encoding capacity to store all their respective locations, and confusion ensues.

However, from a neuroanalytic viewpoint, it is far from clear how a propositional code can be implemented in neural circuitry. The basic causal structure (A > operation > B), where each of the three components has a meaning in terms of stimulus events, does not seem to be readily implementable in a neural network. It is much easier to conceive of parallel template matching as a neural architecture, operating in the form of elaborated receptive field activations on domains of preprocessed information ("feature maps"). For us, the only concept of the operation of propositions is as a temporal playback, where each proposition plays neurally in time in the manner of our recollection of a musical sequence. The logic of the proposition is then encoded as a finite-state

machine, in which each state leads to the most likely subsequent state, and the logic is encoded as the (learned) transitional probabilities. However, this mechanism does not seem to be a plausible basis for letter recognition because the readout of the items does not have a natural sequence or starting point. Moreover, the rapid (10 ms) recognition ability for briefly presented letters (Sperling et al., 1971) does not seem to allow time for any sequential processing.

A second problem with the propositional code is that most people seem to store memories with a significant visuospatial component. If asked for their first recall for the concept of "Queen Elizabeth I", for example, most people will immediately report the visual image of a regal figure resplendent with jewels, which can be inspected for more detail (a tiara in the hair?, a fan in the hand?). The second recall is likely to be of Sir Walter Raleigh gallantly sweeping his cloak into a puddle to save her majesty's shoes. A range of propositions then ensue in some way from these visual image sources: where she was queen, whose daughter she was, how she dealt with the Spanish Armada, and so on. The point is, introspection does not give a clear insight as to how the propositions are coded, although our experience of them is largely sequential. It is clear, however, that we have the capacity for encoding spatial images at some level of detail and accessing them in a spatial manner. Many people have the experience that recalling information from a book includes the location on the page of the word or graph in question. These observations seem to vitiate the idea that information is all stored through a purely propositional code, despite the fact that much cognitive information does have a propositional flavor (such as laws and regulations).

The final problem with the propositional code is that it does not account for the original phenomenology of crowding highlighted in Figures 1 and 2b because there is no basis for perceptual suppression in a propositional code, only for confusion in the identification of the elements.

Attentional tracking

A second model of feature encoding is the attentional tracking concept, derived from the dynamic attentional tracking of Yarbus (1961), Noton and Stark (1971), Pylyshyn (1999), and Intriligator and Cavanagh (2001). The idea is that shape information is encoded by scanpath along the lines or among the features of the scene and that the spatiotemporal form of the scanpath encodes the shape of the pattern. This concept requires no feature processing, although its input could be the outputs of feature-detector arrays (Treisman & Gelade, 1980) or a salience map derived from them (Itti & Koch, 2001). On this attentional scanpath model, the attentional mechanism not only delivers the contents of early visual processing to a conscious processor but also provides a key aspect of the shape encoding in the form of the tracking movements of the attentional spotlight. It seems that this trajectory information would have to be coded as a sequence of operational directions ("down to the right for 1°, up to the left for 1°, back down to the right for 1°, down for 1°" = "Y"). The account of crowding on the scanpath scheme would then derive from the fact that extra items in the tracking sequence would disrupt the sequence and hence impair recognition.

However, in the application to letter recognition, the scanpath coding scheme suffers from the same sorts of problems as the propositional feature/place coding: that the temporal readout requires an invariant start point for the sequence to match, that the coding takes too much time to implement, and that the coding does not have a natural sequence ordering unless it is constrained by an arbitrary rule (e.g., always to move clockwise to the feature with the smallest separation from the immediately preceding location). Measurements of eye-movement scanpaths for scene exploration (e.g., Mannan, Ruddock, & Wooding, 1995; Noton & Stark, 1971; Yarbus, 1961) do not reveal strong consistency in the scanpath sequence, making the sequence a poor vehicle for stable shape encoding. Moreover, the fact that the scanpath trajectory throws away the hard-won local information about orientation, contrast, and color that is provided by the early processing stages makes attentional tracking an inefficient coding option for letter recognition tasks.

Network relaxation minima

The third option for the role of attention in letter recognition is for the local information of elementary feature codes within the attentional spotlight to feed to a recognition network that develops relaxation states corresponding to frequently occurring patterns in the environment, as in the well-known form of the Hopfield net (Hopfield, 1982). This is a neurally plausible implementation of template matching. In the typical analysis, the Hopfield net identifies one of its previously learned stimulus types even when the information is degraded by relaxing the network activation towards the prelearned state corresponding most closely to the input configuration. As specified, this network account of recognition in terms of relaxation states does not address the question of the invariants in the matching (e.g., recognition invariance with respect to size or orientation). We assume that such invariances are somehow solved by parallel processing in the hierarchically scaled network. Computationally, size and orientation invariances are accomplished simply by replicating the input to the recognition algorithm over a range of sizes and orientations. This is a straightforward solution to implementing the invariances that could readily be taken by the neural architecture.

One neural site commonly identified as being involved in recognition is the hippocampus. We know that neurons in the hippocampus and adjacent inferotemporal cortex are capable of highly sophisticated recognition performance

(Fried, Cameron, Yashar, Fong, & Morrow, 2002; Kreiman, Koch, & Fried, 2000; Quiroga, Reddy, Kreiman, Koch, & Fried, 2005),[2] but they should not be expected to shoulder the whole burden of memory storage. In particular, the hippocampus is more likely to be the focus of the attentional comparator operation, drawing in specialized information from relevant cortical regions as needed.

Just as in the previous schemes, the focus of attention in the network relaxation scheme serves to limit the range of information projected to the recognition network and avoid having to analyze the whole visual scene simultaneously. Thus, the relaxation network would play the role of Treisman's "object representation" for the information fed along the attentional pipeline. On this account, we can imagine the relaxation taking place very rapidly because it is operating in parallel across the network for the attended field of features. The key question is how the concept explains the crowding effect. With crowded images, the Hopfield net may still relax to one of its learned classes when extra information is present within the attentional field (or perhaps to an indecisive oscillation between two or more states, in which the output decision is time sensitive). Similar behavior has been described for multilayer network structures (Fukushima, 2005; Takahashi & Kurita, 2005). However, if the distractor information from nearby flanking letters within the attentional focus is too severe, it would distort recognition by violating the parameters of the relevant relaxation states, failing to generate a solution and resulting in an unanalyzed blur such as that in Figure 2b. In some cases, the network may jump into neighboring minima that do not correspond to correct performance, giving rise to the reported confusions.

It is important to emphasize that the recognition network is not an isolated entity within the brain but a node of the whole system that must be connected to the output pathways to communicate the information about the decision. Recognition of a letter inherently implies the ability to say the name of the letter, to have access to its various possible pronunciation sounds, to know its place in the alphabet, to know its role in a plethora of words, and so on. The recognition network must, thus, be intimately connected forward to a range of neural output modules that constitute the functional "meaning" of the letter in its verbal context.

One may ask the question of how the relaxation network is conditioned with the identification information in the first place. This is clearly something encoded in childhood, with the intensive exposure to alphabet letters from the nursery onward. This learning is both bottom–up, in terms of available visual stimuli, and top–down, in terms of parent and teacher insistence that letters are important items with well-established names and roles. However, it is equally clear that such patterns are readily learnable later in life because many people grow up with exposure to one alphabet (say, Korean) and then learn to be proficient readers in another (say, Cyrillic) at college age or later. The computational implementation of a neural network capable of stable recognition of *previously learned* patterns, but rapid development of relaxation states for *new repeated* patterns, has been described by Miyake and Fukushima (1984).

In terms of crowding, it must be the case that the network is trained on letters that are sufficiently isolated to provide a unique signature to generate the requisite relaxation minimum. It seems that these requirements are quite limited, in that the range of elements in most world alphabets use only the elements encodable in a 3×5 grid of pixels, a limit that corresponds to about the resolution of a single V1 hypercolumn (as was pointed out by Schlingensiepen, Campbell, Legge, & Walker, 1986). Measurements of the dimensions of ocular dominance structure in relation to the mapping of human visual cortex suggest that a foveal hypercolumn would analyze a square of about 8 arcmin on a side (Horton & Hocking, 1996; Schira, Wade, & Tyler, 2007), confirming that a single letter is the size of just one hypercolumn near the 20/20 acuity limit (corresponding to about 1 arcmin per letter stroke). The Arabic and Chinese language families exceed this resolution to some degree, but the point still stands in general that an encodable unit is made up of rather few feature elements.

The limited scope of crowding in the fovea at the normal scale for viewing letters (Flom, Weymouth, & Kahneman, 1963) suggests that the encoding phase for the recognition network takes place with foveal viewing of the alphabet items, where the letter is most fully represented. The learning would then transfer to the periphery when attention was induced to access the peripheral location. In this way, the recognition network hypothesis can account for the essentials of the crowding phenomena. Some aspects of this analysis would also be applicable to the other hypotheses, but they do not survive the remaining considerations; thus, the recognition network hypothesis is the preferred one of the three encoding mechanisms considered here.

Conclusion

A full understanding of crowding in letter recognition requires a detailed conceptualization of the process of recognition among large numbers of alternatives. The observed masking properties suggest the operation of recursive inhibition from V3 to V1 as a component of the crowding effect. Although a basic skill, the letter recognition task is sufficiently complex to require mechanisms beyond the level of the known properties of early cortical cells. The goal of the present analysis has been to show how the facts of crowding and other letter recognition phenomena reveal the deficiencies in the existing perceptuo-cognitive models of the underlying processing, which reaches all the way from V1 receptive fields to the operation of conscious attention. The relaxation network concept provides a way forward to understanding

the brain mechanisms involved in letter processing, and the properties of crowding provides interesting insights to constrain our analysis of these mechanisms.

Acknowledgments

This study was supported by NIH/NEI Grant EY 13025 and the Pacific Vision Foundation.

Commercial relationships: none.
Corresponding author: Christopher W. Tyler Ph.D., D.Sc.
Email: cwt@ski.org.
Address: Smith-Kettlewell Eye Research Institute, 2318 Fillmore Street, San Francisco CA 94115, USA.

Footnotes

[1]The omitted text from the third ellipsis only attempts to justify the term "field"; it does not specify its integrative function: "The 'integration field' is just a name for the area circumscribed by the measured critical spacing around the signal (Toet & Levi, 1992). (Levi, Klein & Aitsebamo, 1985, called it the "perceptive field" or "perceptive hypercolumn".) That this area is determined by signal eccentricity, independent of signal and mask size, seems to warrant calling it a 'field'." The term "integration field" is intended to refer to a mechanism in the sense of the term "receptive field", which is generally understood as the mechanism, or response kernel, of a neuron receiving a signal from a sensory surface and generating a response that feeds to a subsequent location in the brain, as specified in the quoted text. The elided text, however, restricts the meaning to a two-dimensional continuum. By defining the "field" as a spatial region of the visual field, the authors seem to be avoiding entirely the question of how information is processed in this spatial region.

[2]Note that the experiments on the specificity of single-neuron responses to high-level features of visual stimuli, such as the responses specific to President Clinton as opposed to other presidents (e.g., Kreiman et al., 2000), are not necessarily incompatible with a network activation concept. Such responses could derive from a network of cells with similar responses that are sparsely sampled by the recording technique.

References

Cannon, M. W., & Fullenkamp, S. C. (1991). Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Research, 31,* 1985–1998. [PubMed]

Cannon, M. W., & Fullenkamp, S. C. (1993). Spatial interactions in apparent contrast: Individual differences in enhancement and suppression effects. *Vision Research, 33,* 1685–1695. [PubMed]

Chen, C. C., & Tyler, C. W. (2001). Lateral sensitivity modulation explains the flanker effect in contrast discrimination. *Proceedings the Royal Society B: Biological Sciences, 268,* 509–516. [PubMed] [Article]

Chen, C. C., & Tyler, C. W. (2002). Lateral modulation of contrast discrimination: Flanker orientation effects. *Journal of Vision, 2*(6):8, 520–530, http://journalofvision.org/2/6/8/, doi:10.1167/2.6.8. [PubMed] [Article]

Chubb, C., Sperling, G., & Solomon, J. A. (1989). Texture interactions determine perceived contrast. *Proceedings of the National Academy of Sciences of the United States of America, 86,* 9631–9635. [PubMed] [Article]

Ejima, Y., & Takahashi, S. (1985). Apparent contrast of a sinusoidal grating in the simultaneous presence of peripheral gratings. *Vision Research, 25,* 1223–1232. [PubMed]

Ellemberg, D., Wilkinson, F., Wilson, H. R., & Arsenault, A. S. (1998). Apparent contrast and spatial frequency of local texture elements. *Journal of the Optical Society of America A, Optics, image science, and vision, 15,* 1733–1739. [PubMed]

Flom, M. C., Weymouth, F. W., & Kahneman, D. (1963). Visual resolution and contour interaction. *Journal of the Optical Society of America, 53,* 1026–1032. [PubMed]

Fried, I., Cameron, K. A., Yashar, S., Fong, R., & Morrow, J. W. (2002). Inhibitory and excitatory responses of single neurons in the human medial temporal lobe during recognition of faces and objects. *Cerebral Cortex, 12,* 575–584. [PubMed] [Article]

Fukushima, K. (2005). Restoring partly occluded patterns: A neural network model. *Neural Networks, 18,* 33–43. [PubMed]

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America, 79,* 2554–2558. [PubMed] [Article]

Horton, J. C., & Hocking, D. R. (1996). Pattern of ocular dominance columns in human striate cortex in strabismic amblyopia. *Visual Neuroscience, 13,* 787–795. [PubMed]

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology, 43,* 171–216. [PubMed]

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews, Neuroscience, 2,* 194–203. [PubMed]

Kontsevich, L. L., & Tyler, C. W. (1999). Distraction of attention and the slope of the psychometric function. *Journal of the Optical Society of America A, Optics, image science, and vision, 16,* 217–222. [PubMed]

Korte, W. (1923). Über die Gestaltauffassung im indirekten Sehen. *Zeitschrift für Psychologie, 93,* 17–82.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience, 3,* 946–953. [PubMed] [Article]

Larsen, A., & Bundesen, C. (1996). A template-matching pandemonium recognizes unconstrained handwritten characters with high accuracy. *Memory & Cognition, 24,* 136–143. [PubMed]

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience, 2,* 375–381. [PubMed] [Article]

Levi, D. M., Klein, S. A., & Aitsebaomo, A. P. (1985). Vernier acuity, crowding and cortical magnification. *Vision Research, 25,* 963–977. [PubMed]

McDonald, J. S., & Tadmor, Y. (2006). The perceived contrast of texture patches embedded in natural images. *Vision Research, 46,* 3098–3104. [PubMed]

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9,* 363–386. [PubMed]

Miyake, S., & Fukushima, K. (1984). A neural network model for the mechanism of feature-extraction. A self-organizing network with feedback inhibition. *Biological Cybernetics, 50,* 377–384. [PubMed]

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science, 171,* 308–311. [PubMed]

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123. [PubMed]

Pelli, D. G., Cavanagh, P., Desimone, R., Tjan, B., & Treisman, A. (2007). Crowding: Including illusory conjunctions, surround suppression, and attention [Abstract]. *Journal of Vision, 7*(2):i, 1, http://journalofvision.org/7/2/i/, doi:10.1167/7.2.i.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12):12, 1136–1169, http://journalofvision.org/4/12/12/, doi:10.1167/4.12.12. [PubMed] [Article]

Põder, E. (2006). Crowding, feature integration, and two kinds of "attention". *Journal of Vision, 6*(2):7, 163–169, http://journalofvision.org/6/2/7/, doi:10.1167/6.2.7. [PubMed] [Article]

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition, 32,* 65–97. [PubMed]

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341–423. [PubMed]

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80,* 127–158. [PubMed]

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435,* 1102–1107. [PubMed]

Schira, M. M., Wade, A. R., & Tyler, C. W. (2007). Two-dimensional mapping of the central and parafoveal visual field to human visual cortex. *Journal of Neurophysiology, 97,* 4284–4295. [PubMed]

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

Snowden, R. J., & Hammett, S. T. (1998). The effects of surround contrast on contrast thresholds, perceived contrast and contrast discrimination. *Vision Research, 38,* 1935–1945. [PubMed]

Sperling, G., Budiansky, J., Spivak, J. G., & Johnson, M. C. (1971). Extremely rapid visual search: The maximum rate of scanning letters for the presence of a numeral. *Science, 174,* 307–311. [PubMed]

Takahashi, T., & Kurita, T. (2005). A robust classifier combined with an auto-associative network for completing partly occluded images. *Neural Networks, 18,* 958–966. [PubMed]

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research, 32,* 1349–1357. [PubMed]

Treisman, A. M. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 353,* 1295–1306. [PubMed] [Article]

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12,* 97–136. [PubMed]

Xing, J., & Heeger, D. J. (2000). Center-surround interactions in foveal and peripheral vision. *Vision Research, 40,* 3065–3072. [PubMed]

Xing, J., & Heeger, D. J. (2001). Measurement and modeling of center–surround suppression and enhancement. *Vision Research, 41,* 571–583. [PubMed]

Yarbus, A. L. (1961). Eye movements during the examination of complicated objects. *Biofizika, 6,* 52–56. [PubMed]

Zeki, S. M. (1978). Uniformity and diversity of structure and function in the monkey prestriate visual cortex. *The Journal of Physiology, 277,* 273–290. [PubMed] [Article]