The nature of letter crowding as revealed by first- and second-order classification images

Anirvan S. Nandy
Department of Psychology, University of Southern California, Los Angeles, CA, USA

Bosco S. Tjan
Department of Psychology and Neuroscience Graduate Program, University of Southern California, Los Angeles, CA, USA

Visual crowding refers to the marked inability to identify an otherwise perfectly identifiable object when it is flanked by other objects. Crowding places a significant limit on form vision in the visual periphery; its mechanism is, however, unknown. Building on the method of signal-clamped classification images (Tjan & Nandy, 2006), we developed a series of first- and second-order classification-image techniques to investigate the nature of crowding without presupposing any model of crowding. Using an “o” versus “x” letter-identification task, we found that (1) crowding significantly reduced the contrast of first-order classification images, although it did not alter the shape of the classification images; (2) response errors during crowding were strongly correlated with the spatial structures of the flankers that resembled those of the erroneously perceived targets; (3) crowding had no systematic effect on intrinsic spatial uncertainty of an observer nor did it suppress feature detection; and (4) analysis of the second-order classification images revealed that crowding reduced the amount of valid features used by the visual system and, at the same time, increased the amount of invalid features used. Our findings strongly support the feature-mislocalization or source-confusion hypothesis as one of the proximal contributors of crowding. Our data also agree with the inappropriate feature-integration account with the requirement that feature integration be a competitive process. However, the feature-masking account and a front-end version of the spatial attention account of crowding are not supported by our data.

Keywords: crowding, letter identification, peripheral vision, classification images


Introduction

Crowding refers to the marked inability in making perceptual judgments when a target object is flanked by other objects (Bouma, 1970; Flom, 1991; Flom, Weymouth, & Kahneman, 1963; Stuart & Burian, 1962; Townsend, Taylor, & Brown, 1971). It is most prominent in peripheral vision and cannot be explained by a lack of spatial resolution. As such, crowding may reveal the critical differences between central and peripheral vision in their mechanisms for perceiving shape. However, the functional and physiological causes of crowding are as yet unsettled.

The attentional accounts of crowding suggest that features from the target and flankers are not bound properly due to a lack of spatial resolution in the attentional mechanism subserving the visual periphery (He, Cavanagh, & Intriligator, 1996; Intriligator & Cavanagh, 2001; Leat, Li, & Epp, 1999; Strasburger, Harvey, & Rentschler, 1991; Tripathy & Cavanagh, 2002). The attentional accounts of crowding are often associated with the claim that crowding originates at a higher level of visual processing. For example, He et al. (1996) showed that a crowded target with indiscernible orientation could produce orientation-specific adaptation, suggesting that crowding occurs at a stage beyond the primary visual cortex. However, using a similar paradigm, Blake, Tadin, Sobel, Raisian, and Chong (2006) showed that the strength of orientation-specific adaptation could be modulated by the severity of crowding when the contrast of the adapting pattern was kept low. Contrary to He et al., this later finding argues for a low-level origin of crowding while being ambivalent about the attentional account in general.

Arguing against the attentional accounts, Pelli, Palomares, and Majaj (2004) stated that a parsimonious explanation of crowding need not involve attention. They went as far as suggesting that phenomena that were generally associated with attention, such as illusory conjunction (Treisman & Schmidt, 1982), shared the same mechanistic root as crowding in lower level visual processing. This view implies an intriguing possibility that finding the mechanism(s) of crowding will also resolve a broader set of issues concerning attention.

Regardless of whether the root cause of crowding involves attention, most theories propose some form of interaction between simple features as a proximal cause of crowding. At least three forms of interaction have been proposed: masking, inappropriate feature integration, and source confusion or feature mislocalization.
A masking account of crowding argues that the sensitivity to the simple features constituting the target is suppressed by the presence of the flankers. Chung, Levi, and Legge (2001) identified the similarity and distinctiveness between ordinary masking (target and masker spatially overlap) and crowding (target and masker do not overlap). Chung et al. suggested that crowding and ordinary masking share the same linear filtering stage early in visual processing but differ in how responses are pooled spatially in the later stages. Later studies emphasized the difference between crowding and ordinary masking. A critical difference is that whereas ordinary masking affects both detection and identification of visual patterns (Thomas, 1985), crowding in the periphery affects only identification (Levi, Hariharan, & Klein, 2002; Pelli et al., 2004). Another masking account of crowding is surround suppression. Surround suppression refers to the physiological finding of a reduction in a neuron’s response to an otherwise optimal stimulus when certain patterns are presented outside the neuron’s “classical” receptive field (Allman, Miezin, & McGuinness, 1985; Hubel & Wiesel, 1965). Petrov and McKee (2006) showed a number of similarities between crowding and surround suppression, including the anisotropy of the spatial interaction zones observed in crowding (Toet & Levi, 1992). Although the current consensus is that crowding cannot be explained by ordinary masking, surround suppression as a mechanism for crowding remains viable.

That crowding in the periphery affects identification but not detection suggested a dissociation between the process of detecting simple features and the subsequent process of integrating these features into a recognizable pattern (Levi et al., 2002; Pelli et al., 2004). An inappropriate integration account of crowding postulates that the deficiencies of feature integration in the periphery lead to crowding. Levi et al. (2002) suggested that the flaw of feature integration originates from the use of a defective template that is not well matched to the target at a processing stage beyond feature detection. Parkes, Lund, Angelucci, Solomon, and Morgan (2001) argued that feature attributes such as orientation appear to be compulsorily averaged such that although the mean can be accurately recorded, the individual feature attributes are inaccessible. A general form of this claim is that feature integration in the periphery amounts to computing group statistics of the various feature attributes within a spatial region. While these group statistics are accessible, the individual sample values are lost.

A source-confusion account of crowding suggests that features from the flankers are mistaken to be features of the target (Krumhansl & Thomas, 1977; Wolford, 1975). While this account is often associated with the attentional account of crowding (e.g., Strasburger, 2005; Strasburger et al., 1991), it is equally consistent with the fact that spatial uncertainty in the periphery is high (Levi & Klein, 1986; Levi, Klein, & Yap, 1987; Pelli, 1985).

It is to be noted that none of the accounts of crowding are mutually exclusive. The goal of this study was to use a data-driven approach to elucidate the mechanisms of letter crowding without making a priori assumptions about the possible mechanisms. We will stay mostly neutral as to whether attention is a root cause of crowding. However, our results will address the three proximal accounts of crowding: masking, inappropriate feature integration, and source confusion.

Building on our method of signal-clamped classification images (Tjan & Nandy, 2006), we developed three novel analytic procedures to reveal the features and their spatial configurations used by the central and peripheral visual systems in a letter-identification task when the target letter was either presented alone or flanked by other letters.

Classification-image methods (Ahumada 2002; Beard & Ahumada, 1999) have been very useful in revealing an observer’s visual strategy (e.g., vernier acuity: Beard & Ahumada, 1999; stereopsis: Neri, Parker, & Blakemore, 1999; illusory-contour perception: Gold, Murray, Bennett, & Sekuler, 2000; identification of facial expression: Adolphs et al., 2005; Gosselin & Schyns, 2003; surround effect on contrast discrimination: Shimozaki, Eckstein, & Abbey, 2005). By “visual strategy,” we mean the parts of the stimulus that are used by a subject to perform the task. We will loosely refer to such a visual strategy revealed by a classification-image method as a “template” and take the term to mean the spatiotemporal average of the features used by the visual system. We recently showed that the standard classification-image technique can be easily extended to overcome the spatial uncertainty that is either present in the stimuli or intrinsic to the visual system, thereby revealing the shift-invariant template used by the visual system (Tjan & Nandy, 2006). Because the human periphery is known to exhibit a significant amount of intrinsic spatial uncertainty (Levi & Klein, 1986; Levi et al., 1987; Pelli, 1985), our signal-clamping method affords us the ability to examine the perceptual templates that are used with and without letter crowding. This helps to address one of the fundamental questions regarding the nature of crowding: Is crowding caused by a distortion in the perceptual template?

To address the issue of source confusion or feature mislocalization, we developed a method to infer the specific structural characteristics of the flankers that correlate with crowding-induced errors. The results allow us to posit the possible contributors to crowding.

To address the issues related to feature detection and integration, we showed that the noise fields in signal-clamped classification images contain sufficient information to reveal the second-order correlation structures of subtemplate features. Computing these correlation structures allows us to infer the possible shape of the putative features, compare them to features used by an ideal-observer model, and determine the spatial region from which these features were extracted by the visual system.

In summary, we probed into the nature of letter crowding by estimating and comparing the following between
crowded and noncrowded conditions: (1) first-order classification images, which are the spatiotemporal average of features; (2) structural effect of the flanking letters; (3) spatial extent of feature utilization; and (4) second-order feature maps. To preview, we found that (1) crowding did not cause distortions in the perceptual templates, (2) response errors during crowding were strongly correlated with spatial structures of the flankers, (3) intrinsic spatial uncertainty was not systematically affected by crowding, and (4) crowding reduced the amount of valid features utilized by the visual system and, at the same time, increased the amount of invalid features used.

**Analysis procedures**

In this section, we will describe three novel procedures for the analysis of crowding: (1) a procedure to calculate and visualize any systematic structures in the flankers that lead to errors under crowding, (2) a procedure to estimate the spatial region over which features are detected and utilized, and (3) a procedure to calculate and visualize the features in terms of their second-order statistics. Of these three procedures, the first is applicable only to classification-image experiments with flankers present. The last two procedures are broader in scope and can be applied to any signal-clamped classification-image data. In the next few paragraphs, we will provide a brief conceptual overview of the analytical procedures. Details of the procedures will be given in the subsections that follow.

As we shall describe in greater detail in the Methods section, the general setup of the experiment was that of letter identification in noise. Briefly, in different blocks, the target letter ("x" or "o") was either presented alone or was flanked by other letters. Both the target and the flankers were masked by Gaussian luminance noise with a constant spectral density. The contrast of the target and the flankers was the same and was adjusted to maintain an accuracy level of 75% correct. We tested at the fovea and in the periphery.

Obtaining classification images in the periphery can be challenging because of the high intrinsic spatial uncertainty in the periphery. We were able to significantly reduce the effect of uncertainty by presenting a relatively high contrast signal in our classification-image experiments. This "signal-clamping" technique was developed in Tjan and Nandy (2006), which provides a detailed analysis of its various properties. A brief intuitive exposition is provided in the Signal-clamped classification images section.

To assess and visualize the effect of the flankers on an observer's perception of the target letter, we extended the first-order classification-image analysis to include the flanking letters as part of the masking noise. We constructed the classification images by testing for deviations from the expected pixel-wise mean of this composite noise when the flankers and the masking noise were sorted according to the observer's response for each presented target letter (see the Flanker analysis section). The result is a map that reveals the structural elements of the flankers that influenced the observer's response, thereby allowing a direct assessment of source confusion of visual features under crowding.

A methodological point of departure between the current study and most of the early work involving classification images is that we were able to reveal properties of the subtemplate features used by the visual system beyond their spatiotemporal averages (which are the conventional classification images). For each trial, we computed the correlations between pairs of noise pixels as a function of their relative separations in space. Because the masking noise was white, the expected pairwise correlation would be zero. Any accidental configuration of noise pixels that resemble a feature used by the visual system would systematically affect the observer's response. Thus, any nonzero correlations that emerge as a result of the analysis reveal the second-order correlational structure of the subtemplate features (see the Feature maps section). Comparing human feature maps with those obtained from an ideal-observer model provides an assessment of the utilization and validity of the features used by the visual system under different experimental conditions.

A prerequisite for computing the pairwise correlations is to decide on a region in the noise field for carrying out the computation. If the region is too small, the correlations will be weak due to insufficient sample size. If it is too large, it will include regions of the noise ignored by the visual system, thereby diluting the correlations that were due to the features used. We developed a method to search for a region in the stimulus that leads to a maximum level of correlation (see the Optimal region of interest section). An important by-product of finding this optimal region of interest (ROI_opt) is that it reveals the region in the image where task-relevant features were extracted by the visual system.

**Signal-clamped classification images**

Signal-clamped classification images (Tjan & Nandy, 2006) are simply the classification images obtained at high signal contrast (which serves to "clamp down" a particular perceptual channel, thus minimizing the effect of uncertainty) and without adding up the subimages from different stimulus–response categories. The analysis of signal-clamped classification images focuses on the data from the error trials. As shown in Tjan and Nandy (2006), a unique property of signal-clamped classification images is that any spatial uncertainty intrinsic to the observer is significantly reduced or even eliminated for the "miss"
component of an error trial. For conditions in which there is a significant amount of intrinsic spatial uncertainty in the visual system (e.g., form vision in the periphery), this property of the signal-clamped classification-image method dissociates the miss components from the false-alarm components of an error trial and allows separate imaging of the shift-invariant perceptual template for each of the stimuli. In this study, we began with experimental conditions suitable for extracting signal-clamped classification images and proceeded to analyze only the error trials. The detailed experimental setup will be given in the Methods section.

In this study, we began with experimental conditions suitable for extracting signal-clamped classification images and proceeded to analyze only the error trials. The detailed experimental setup will be given in the Methods section. The analytic and empirical properties of signal-clamped classification images in general are given in the study of Tjan and Nandy. For the purpose of this article, it is sufficient to bear in mind that an error-trial subimage contains the following: (1) a perceptual template that is negatively correlated with the clamped target mechanism and (2) a spatially dispersed “haze” that corresponds to the unclamped alternative mechanism.

**Flanker analysis**

In conditions when flankers were present and influenced performance, we want to determine if specific parts from the flankers affected an observer’s response, leading to crowding, and if these parts resembled those of the targets. The procedure is similar to that used to obtain the standard classification subimages with a few modifications. First, instead of classifying only the noise field at each trial, the sum of the noise and the flanker images are classified, according to the stimulus presented and the observer’s response, and then averaged. We thus obtain classification subimages of the following form, exemplified here for the case in which the target presented was “o” while the response was “x”:

\[
I_{OX} = \frac{1}{n_{OX}} \sum_i \left( N_{OX,i} + c_{OX,i} F_{OX,i} \right),
\]

where \( N_{OX,i} \) is the noise field from trial \( i \), \( F_{OX,i} \) is the flanker image at unit contrast and \( c_{OX,i} \) is the flanker (and target) contrast for the trial, and \( n_{OX} \) is the number of trials in that category. This classification subimage represents the category mean for the population of trials in which the presented target was “o” and the response was “x”.

Similarly, we can obtain a classification subimage for all the trials in which the presented target was “o”, irrespective of the response. Let us call this \( I_{O*} \). This subimage represents the expected mean for the population of trials in which the target was “o”, regardless of the behavioral response. As one further step, this expected mean (under the null hypothesis that flankers have no affect on response) needs to be corrected for contrast to account for the slight difference in contrast between the error trials and the correct trials. The difference exists because (1) the contrast of the target and the flankers were varied dynamically from trial to trial using QUEST (see the Methods section) and (2) the probability of a subject making a discrimination error is higher when the trial contrast is slightly below the average population contrast. Thus, the average contrast for the error trials is slightly lower than the average population contrast. If we do not correct for this difference, it will lead to a biased estimate of the \( z \) scores. The correction can be expressed as follows.

\[
\tau_{OX} = \frac{1}{n_{OX}} \sum_i c_{OX,i}
\]

\[
r = \tau_{OX} - \tau_{O*}
\]

\[
I_{O*} = I_{O*} + \frac{r}{n_{O*}} \sum_i c_{O*,i} F_{O*,i}
\]

where \( \tau_{OX} \) is the average contrast for the target “o”–response “x” trials, and \( \tau_{O*} \) is the average contrast for all \( n_{O*} \) target “o” trials (irrespective of the response). Thus, \( I_{O*} \) represents the contrast-corrected version of \( I_{O*} \).

Finally, under assumptions of normality, because \( n_{OX} \) is typically large, we can perform a z test between the category mean and the expected mean under the null hypothesis and obtain a z-score map as follows:

\[
SS_{I_{O*}} = \sum_i \left( N_{O*,i} + c_{O*,i} F_{O*,i} \right)^2
\]

\[
\sigma_{I_{O*}} = \sqrt{SS_{I_{O*}} - n_{O*} \tau_{O*}^2} / n_{O*}
\]

\[
Z_{OX} = \frac{I_{OX} - I_{O*}}{\sigma_{I_{O*}} \sqrt{n_{O*} / n_{OX}}}
\]

where \( \sigma_{I_{O*}} \) is the standard error of the expected mean under the null hypothesis.

Under the null hypothesis that there is no effect of the flankers, we would not expect to see any significant \( z \) scores in the regions where the flankers were presented. On the other hand, if we do find significant \( z \) scores in these regions, then displaying the \( z \)-score map (\( Z_{O*} \)) as a color-coded image provides a way to visualize the presence (indicated by statistically significant positive \( z \) scores) or the absence (negative \( z \) scores) of pixels in the flankers (and elsewhere) that bias the observer toward responding “x”. Similarly, we can obtain the \( z \)-score map, \( Z_{XO} \), for the “o” response. This procedure is illustrated in Figure 1.
Another way of visualizing the same information is to perform a $t$ test between the two error response categories:

$$
\bar{F}_{OX} = \frac{1}{n_{OX}} \sum_i \left( N_{OX,i} + F_{OX,i} \cdot c_{OX,i} \right)^2
$$

$$
\sigma_{pool} = \sqrt{\frac{n_{OX} (\bar{F}_{OX} - T_{OX}^2) + n_{XO} (\bar{F}_{XO} - T_{XO}^2)}{n_{OX} + n_{XO} - 2}}
$$

$$
T = \frac{T_{XO} - T_{OX}}{\sigma_{pool} \sqrt{\frac{1}{n_{OX}} + \frac{1}{n_{XO}}}}.
$$

The $t$-test map, $T$, provides a direct contrast of the pixels (significant positive $t$ values) that bias toward the “$o$” response versus the features (significant negative $t$ values) that bias toward the “$x$” response.

### Feature maps

For our “$o$” versus “$x$” task, we can postulate two competing mechanisms, one selective to “$o$” and the other selective to “$x$”. The error-trial noise fields of signal-clamped classification images provide a natural way of separating these two mechanisms. This is because the high-contrast target reduces the effect of intrinsic positional uncertainty only in the mechanism selective to the presented target. Briefly, an error response occurs when the “target” mechanism (the one selective to the presented target) misses or the “alternative” mechanism (the one selective to the target not presented) false-alarms or both. In Tjan and Nandy (2006), we showed that by presenting a high-contrast target at a fixed location, one of the many spatially dispersed channels in the target mechanism can be localized, in the sense that it leads to most of the responses. When a miss occurs in this mechanism, it is most likely due to the fact that the masking noise suppresses this one channel that would otherwise be “clamped” to the presented target. As a result, the average of the error-trial noise fields yields a clear template that is almost unaffected by intrinsic spatial uncertainty and is negatively correlated with the template used by the target mechanism. In contrast, none of the channels in the alternative mechanism are clamped or biased by the presented signal. When an error is caused by a false alarm from this alternative mechanism, the false alarm can originate from any of its spatially dispersed channels. As a result, the average of the noise fields does not contain any clear template for this mech-
anism if the intrinsic spatial uncertainty in the mechanism is high. In sum, in the average (thus, first-order) noise field from the error trials, the template for the mechanism that missed is revealed but the template for the mechanism that false-alarmed is not.

If we remove, by projection, the spatially localized first-order template (say, for the “o” mechanism) from the corresponding error-trial noise samples \((N_{OX,i})\), the residual noise samples will contain mostly the spatially dispersed structures corresponding to the alternative (“x”) mechanism. Extracting the second-order statistics from the residual noise fields can therefore reveal the spatial structure of the sub-template fragments (“features”) of the alternative (“x”) mechanism. Similarly, we can obtain the second-order feature statistics for the “o” mechanism from the error-trial noise samples \(N_{XO}\). In other words, the first-order statistics of the noise fields \(N_{OX}\) give us the shift-invariant template of the “o” mechanism, which represents the spatiotemporal structures corresponding to the alternative (“x”) mechanism. Similarly, we can obtain the second-order feature statistics for the “o” mechanism from the error-trial noise fields \(N_{XO}\) (note the swapping of the subscripts) reveal the sub-template features.

Next, we describe the steps to extract the second-order feature statistics.

As indicated above, we first project out the first-order signal-clamped template from the error-trial noise fields. This is shown below for the case in which the clamped stimulus was “o” and the response was “x”, \(N_{OX,i}^{d}\) is a down-sampled version of the noise field and \(N_{OX}\) is the average down-sampled noise field, which contains the clamped template for the “o” mechanism to be removed by projection. In the above equation, \(N_{OX,i}^{d}\) and \(\overline{N}_{OX,i}^{d}\) represent the vectorized form of the corresponding noise fields. Removal by projection amounts to subtracting out the vector component in the direction of the vector representing the first-order template. The purpose of the down-sampling operation is to preclude any spurious correlations due to the stroke widths of the letter stimuli. For example, if a particular stroke of the letter “x” is 3 pixels wide, this will introduce correlation in the error-trial noise pixels that are congruent with the stroke width. Down-sampling the noise field by a factor of 3 will reduce the effect of this correlation in the final feature maps. A letter in a down-sampled image would have a stroke width of 1 pixel. This is accomplished by first convolving the noise field with an \(f\)-by-\(f\) rectangular mask \(M\) that preserves the signal-to-noise ratio (SNR) followed by a down-sampling operation by a linear factor \(f\). The down-sampling factor \(f\) is chosen such that it matches the mean stroke width among the presented target letters. For the foveal viewing conditions, the down-sampling operation is not performed because the letter stimuli have a stroke width of 1.

Next, we identify an optimal region-of-interest (ROI\(_{opt}\)) in the residual noise fields \(N_{OX,i}\) for extracting the second-order statistics (see Equation 14); this critical step of identifying the optimal ROI will be discussed in the next section. We then calculate, in the standard fashion, the pixel-wise correlation coefficient between all points in ROI\(_{opt}\) and another region ROI\(_{off}\) within the same noise field that is offset from ROI\(_{opt}\) by \((\Delta x, \Delta y)\) and accumulate the correlation coefficient across all the residual noise fields, \(N_{OX,i}\):

\[
\begin{align*}
M &= \frac{1}{f^2} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{f \times f}, \\
N_{OX,i}^{d}(x,y) &= [N_{OX,i} \otimes M](xf, yf) \\
\overline{N}_{OX,i} &= \overline{N}_{OX,i}^{d} = \left(\overline{N}_{OX,i}^{d} \cdot \overline{N}_{OX}\right) \overline{N}_{OX}^{d} \\
\text{where } N_{OX,i} &\text{ is the noise field for the } i\text{th trial when the stimulus was “o” and the response was “x”, }N_{OX,i}^{d} \text{ is a down-sampled version of the noise field and } \overline{N}_{OX} \text{ is the average down-sampled noise field, which contains the clamped template for the “o” mechanism to be removed by projection. In the above equation, } \overline{N}_{OX,i}^{d} \text{ and } \overline{N}_{OX,i}^{d} \text{ represent the vectorized form of the corresponding noise fields. Removal by projection amounts to subtracting out the vector component in the direction of the vector representing the first-order template. The purpose of the down-sampling operation is to preclude any spurious correlations due to the stroke widths of the letter stimuli. For example, if a particular stroke of the letter “x” is 3 pixels wide, this will introduce correlation in the error-trial noise pixels that are congruent with the stroke width. Down-sampling the noise field by a factor of 3 will reduce the effect of this correlation in the final feature maps. A letter in a down-sampled image would have a stroke width of 1 pixel. This is accomplished by first convolving the noise field with an } f\text{-by-} f \text{ rectangular mask } M \text{ that preserves the signal-to-noise ratio (SNR) followed by a down-sampling operation by a linear factor } f. \text{ The down-sampling factor } f \text{ is chosen such that it matches the mean stroke width among the presented target letters. For the foveal viewing conditions, the down-sampling operation is not performed because the letter stimuli have a stroke width of 1. Next, we identify an optimal region-of-interest (ROI}_{opt}) \text{ in the residual noise fields } N_{OX,i} \text{ for extracting the second-order statistics (see Equation 14); this critical step of identifying the optimal ROI will be discussed in the next section. We then calculate, in the standard fashion, the pixel-wise correlation coefficient between all points in ROI}_{opt} \text{ and another region ROI}_{off} \text{ within the same noise field that is offset from ROI}_{opt} \text{ by } (\Delta x, \Delta y) \text{ and accumulate the correlation coefficient across all the residual noise fields, } N_{OX,i}:}
\end{align*}
\]

\[
\begin{align*}
\mu_{ROI_{opt}} &= \frac{1}{n_{ROI_{opt}}} \sum_{i=1}^{n_{ROI_{opt}}} \sum_{j=1, (j,k) \in ROI_{opt}} N_{OX,i}(j,k) \\
\mu_{ROI_{off}} &= \frac{1}{n_{ROI_{off}}} \sum_{i=1}^{n_{ROI_{off}}} \sum_{j=1, (j,k) \in ROI_{opt}} N_{OX,i}(j+\Delta x, k+\Delta y) \\
SS_{ROI_{opt}} &= \sum_{i=1, (j,k) \in ROI_{opt}} \sum_{i=1}^{n_{ROI_{opt}}} \left( N_{OX,i}(j,k) - \mu_{ROI_{opt}} \right) \\
SS_{ROI_{off}} &= \sum_{i=1, (j,k) \in ROI_{opt}} \sum_{i=1}^{n_{ROI_{off}}} \left( N_{OX,i}(j+\Delta x, k+\Delta y) - \mu_{ROI_{off}} \right) \\
SS_{ROI_{opt},ROI_{off}} &= \sum_{i=1, (j,k) \in ROI_{opt}} \sum_{i=1}^{n_{ROI_{opt}}} \left[ \left( N_{OX,i}(j+\Delta x, k+\Delta y) - \mu_{ROI_{off}} \right) \times \left( N_{OX,i}(j,k) - \mu_{ROI_{opt}} \right) \right] \\
r_{OXY}(\Delta x, \Delta y) &= \frac{SS_{ROI_{opt},ROI_{off}}}{\sqrt{SS_{ROI_{opt}} \cdot SS_{ROI_{off}}}}.
\end{align*}
\]

This correlation coefficient, \(r_{OXY}(\Delta x, \Delta y)\), is evaluated for a range of displacements \((\Delta x, \Delta y)\) within ±1 letter size measured in units of the height of the letter “x” (i.e., \(-L \leq \Delta x, \Delta y \leq L\), where \(L\) is the x-height in pixels). To aid direct comparisons across subjects and conditions with different sizes of ROI_{opt} and different numbers of error trials, we convert the correlation coefficient to Fisher’s Z.
(Papoulis, 1990), such that the correlation at a displacement \((\Delta x, \Delta y)\) is expressed in terms of a random variable \(r_{Z_{OX}}(\Delta x, \Delta y)\) with an approximate standard normal distribution:

\[
r_{Z_{OX}}(\Delta x, \Delta y) = \frac{1}{2} \ln \left( \frac{1 + r_{OX}(\Delta x, \Delta y)}{1 - r_{OX}(\Delta x, \Delta y)} \right) \cdot \sqrt{n_{OX}} - 3.
\]  

(7)

The ensemble of correlation coefficients (expressed in terms of Fisher’s Z) can then be plotted on a color-coded map \(r_{Z_{OX}}\). The map (see Figure 2) thus obtained can be interpreted as follows: The center of the map represents any pixel in ROI\(_{opt}\); a “hot” spot or a positive correlation at an offset \((\Delta x, \Delta y)\) from the center indicates that the observer is biased toward a particular response when two pixels separated by \((\Delta x, \Delta y)\) have the same contrast polarity; a “cold” spot or a negative correlation at an offset \((\Delta x, \Delta y)\) from the center indicates that the response is partly driven by two pixels separated by \((\Delta x, \Delta y)\) with opposite contrast polarity.

By comparing feature maps obtained from an ideal-observer model\(^2\) (for which we know the ground truth) to that of a human observer, one can uncover the nature of the strategy employed by the visual system at an atomic level that has not been possible with standard classification images. Figure 9 shows the ideal-observer feature maps obtained from stimuli used in our study. The ideal-observer model, detailed in Tjan and Nandy (2006), was limited by static white contrast noise and an intrinsic spatial uncertainty equated to \(\pm 1.5\) letter size (x-height) in

Figure 2. Procedure to calculate the second-order feature maps. Correlation coefficients between each pixel in an ROI\(_{opt}\) and the corresponding pixel in an offset region (ROI\(_{off}\)) are accumulated over all residual noise fields for a particular error-response category and are plotted in terms of Fisher’s Z to facilitate comparison across conditions and subjects on a color-coded map. This map reveals the second-order statistics of the features that comprise an observer’s perceptual template. The center of the map represents any pixel in ROI\(_{opt}\); a “hot” spot or a positive correlation at an offset \((\Delta x, \Delta y)\) from the center indicates that the observer is biased toward a particular response when two pixels separated by \((\Delta x, \Delta y)\) have the same contrast polarity; a “cold” spot or a negative correlation at an offset \((\Delta x, \Delta y)\) from the center indicates that the response is partly driven by two pixels separated by \((\Delta x, \Delta y)\) with opposite contrast polarity.
both horizontal and vertical directions. For our purpose, the ideal-observer model was given the task of discriminating the letter “o” from “x” using the same letter images shown to our subjects but in the absence of flankers. The level of the white noise is inconsequential and was set to an rms contrast of 1.0. Figure 10 in the Results and discussions section shows the human feature maps. Many challenges abound in trying to quantitatively compare the model and human data. These include the following: (1) the correlation coefficients and the corresponding Fisher’s Z values are not directly comparable because of the differences in internal noise and uncertainty between model and human observers; (2) the representation used by the human observers may correspond to letters of a slightly different shape and size than the ones used in the experiments, resulting in second-order correlations that are morphologically similar to features used by the ideal-observer model but are spatially displaced; (3) the process of estimating the correlation coefficients from human data is noisy. We overcame these challenges by considering thresholded versions of the correlation maps to extract three quantities (separately for the two putative “o” and “x” mechanisms): (1) quality of match \(Q_m\) between human and ideal-observer feature maps, (2) proportion of ideal-observer features used by the human observer \(U\), and (3) proportion of features used by the human observer that are valid according to the ideal-observer model \(V\).

Because each of the target letters was essentially a binary image, with background at zero contrast and foreground (letter) at a positive contrast, the second-order features directly associated with a letter always has a positive correlation coefficient. The negative correlation in the ideal-observer (and human) feature maps originated from the difference between the two letters in the presence of spatial uncertainty. Therefore, in our analysis, we consider only the positively correlated features. To further simplify the analysis, we consider only the presence or absence of a feature, discarding the magnitude of the correlation coefficients by thresholding the second-order feature maps.

The basic problem of comparing the human observer feature map to that of an ideal-observer feature map lies in the arbitrariness of setting thresholds for the maps to decide what counts as a feature. This problem can be overcome by performing the relevant computations at all possible threshold settings, akin to tracing out receiver operating characteristic (ROC) curves.

Thus, for the purpose of calculating \(Q_m\) (e.g., for the “o” mechanism), we evaluate a family of ROC curves parameterized by a threshold \(\lambda_{hm}\) for defining the level of correlation that may constitute a feature used by a human observer. Each ROC curve is in turn defined by a set of points \((h_{O,\lambda_{hm}}(\lambda), f_{O,\lambda_{hm}}(\lambda))\) depending on a threshold \(\lambda\) for demarcating the ideal-observer features (a “hit” \([h]\) corresponds to the presence of an ideal-observer feature at threshold \(\lambda\) when the same feature is also present in the human feature map; similarly, a “false-alarm” \([f]\) corresponds to the presence of an ideal-observer feature without the presence of the same feature in the human map):

\[
h_{O,\lambda_{hm}}(\lambda) = \Pr(rZ^\text{ideal}_O > \lambda | rZ^\text{hm}_O > \lambda_{hm})
\]

\[
f_{O,\lambda_{hm}}(\lambda) = \Pr(rZ^\text{ideal}_O > \lambda | rZ^\text{hm}_O \leq \lambda_{hm})
\]

with \(rZ^\text{ideal}_O\) and \(rZ^\text{hm}_O\) being the second-order feature maps for the “o” mechanism obtained from the ideal-observer model and a human observer, respectively. \(\lambda_{hm}\) is chosen to span the positive range of \(rZ\) values in the human map. For the ideal-observer feature map, the threshold \(\lambda\) spans the full range of the \(rZ\) values comprising the map, thereby tracing out the ROC curve. We define the quality of match between the model and human feature map as the area under the ROC curve (AUC) maximized with respect to \(\lambda_{hm}\):

\[
\text{AUC}_O(\lambda_{hm}) = \int_{\lambda_{hm}}^{\infty} h_{O,\lambda_{hm}}(\lambda) d(f_{O,\lambda_{hm}}(\lambda))
\]

\[
Q_{m,O} = \max_{\lambda_{hm}} \text{AUC}_O(\lambda_{hm}).
\]

Note that \(Q_{m,O}\) does not depend on any arbitrary thresholds for either the ideal-observer model or the human. It is also insensitive to the raw amplitude of correlation coefficients of either human or ideal-observer feature maps and is robust with respect to the level of an observer’s intrinsic noise, provided that the number of trials used to estimate the feature maps is sufficient.

The ideal-observer model uses the entire letter as a unitary template “feature”. Its second-order feature map therefore contains all pairwise correlations between pixels on a letter (excluding correlation of equal strength on both letters). If the human visual system uses a combination of subtemplate fragments as features to identify the letters (e.g., using a pair of vertical parallel lines to identify the letter “o”), the human feature map, in terms of its regions with positive correlation coefficients, will be a subset of the ideal-observer feature map. A corollary of this observation is that any human feature that is not present in the ideal-observer feature map is a spurious feature, one that may lead to erroneous performance. Therefore, we are interested in quantifying the proportion of ideal-observer features that are also used by a human observer \(U\), as well as the proportion of human features that are valid features \(V\).

To estimate these two quantities, we first set a threshold of \(rZ = 1\), which, in terms of Fisher’s Z, corresponds to 1 SD away from zero correlation in both the human and ideal-observer maps; that is, we consider a feature as any correlation with Fisher’s Z greater than 1.0. The qualitative results are not sensitive to this somewhat arbitrary threshold. We define the proportion of ideal-observer features used by a human observer \(U\), or “feature utilization”) as

\[
U_O = \Pr(rZ^\text{hm}_O > 1 | rZ^\text{ideal}_O > 1).
\]
We define the proportion of human features used by the ideal-observer model \((V,\) or “feature validity”) as

\[ V_O = h_{O,1}(1), \]

where \(h\) is defined in Equation 8. Standard error estimates for the three quantities \(Q_m, U,\) and \(V\) are obtained by bootstrapping on the human feature maps (Efron & Tibshirani, 1993).

**Optimal region of interest**

Let us return to the estimation of the second-order feature maps \((rZ)\). With one critical exception, our method of computing the correlation coefficients is mathematically analogous to averaging the power spectra of the noise sample within each stimulus–response category, which have been applied with stimuli of narrow spatial bandwidth (e.g., Gabors) and large spatial extent relative to the noise field (Solomon, 2002; Thomas & Knoblauch 2005). A significant difficulty of applying the same technique to spatially broadband stimuli such as letters is that while the stimuli are highly localized in space, there can be a significant amount of spatial uncertainty intrinsic to the human observer. Computing the spectrum of the entire noise field will render any highly localized correlation undetectable. However, because of intrinsic spatial uncertainty, it is also unwise to compute the spectrum (or equivalently, pairwise correlation) only in the target region. The approach we took was to search for an ROI within the noise field that maximized a measure of pairwise correlation. This ROI reveals the spatial range within which features are extracted by the observer.

The intuition is as follows (see Figure 3): Consider a noise field with a central correlated region; if the ROI chosen to calculate the correlation coefficient is small, then the level of significance will be low (or equivalently, with a high \(p\) value) because the ROI is insufficient to capture all the correlations in the data; the level of significance will increase as the size of the ROI is increased; if, however, the ROI is too large such that it includes mostly the uncorrelated noise, the level of significance again drops. Thus, the optimal ROI is the one with the maximum level of significance (or alternatively, the minimum \(p\) value).

For computational tractability, we restrict our candidate ROIs to be rectangular and centered in a residual noise field (Equation 5) to coincide with the center of the target letter. First, an ROI of size \(h \times w\) pixels \((R_{h,w})\) is selected from the set

\[ h, w \in \{0.5L, 1.0L, 1.5L, 2.0L, 2.5L, 3.0L\}, \]

where \(L\) is the letter size (x-height) in pixels. The significance level of this ROI is then calculated as the mean of \(log\) \(p\) values:

\[ P_{h,w} = \frac{1}{L^2} \sum_{\Delta x} \sum_{\Delta y} \log(p_{\Delta x, \Delta y}), \]

where the range of \(\Delta x\) and \(\Delta y\) is restricted to within \(\pm 1\) letter size (i.e., \(-L \leq \Delta x, \Delta y \leq L\)), where \(L\) is the x-height in pixels, and \(p_{\Delta x, \Delta y}\) is the \(p\) value of the correlation coefficient \(r(\Delta x, \Delta y)\) between ROI_\(h,w\) and a region that is offset from ROI_\(h,w\) by \((\Delta x, \Delta y)\) as defined in Equation 6. \(P_{h,w}\) is calculated for all possible combinations of \(h\) and \(w\) chosen from the set given in Equation 12. Finally, the optimal ROI is selected as follows:

\[ (h_{\text{min}}, w_{\text{min}}) = \arg\min_{h,w} P_{h,w}, \]

\[ \text{ROI}_{\text{opt}} = \text{ROI}_{h_{\text{min}}, w_{\text{min}}}. \]

In essence, the optimal ROI is the ROI that minimizes the geometric mean of \(p\) values in the resulting correlation.
map according to Equation 6. The optimal ROI defined in Equation 14 is used to calculate the second-order feature maps described in the previous section. It also provides an estimate of the spatial extent over which features are detected and utilized. In this light, we will refer to the optimal ROI as the feature utilization zone and we will compare the feature utilization zones between different experimental conditions (flanked vs. unflanked; foveal vs. peripheral).

**Methods**

Human experiments were conducted with four different experimental conditions. There were two viewing conditions—foveal and peripheral at 10° in the inferior visual field. In each of the viewing conditions, lowercase target letters “o” and “x” were presented either singly (unflanked) or were flanked on either side by two other lowercase letters. The foveal conditions (flanked and unflanked) served primarily as control conditions to enable direct (within subject) comparison with the corresponding peripheral conditions.

**Subjects**

Three subjects (BB, LM, and OR, all students at the University of Southern California) with normal or corrected-to-normal vision and naïve to the purpose of the study participated in the experiments. All had (corrected) acuity of 20/20 or better in both eyes. Subjects viewed the stimuli binocularly in a dark room with a dim night-light. Written informed consent was obtained from each subject before the commencement of data collection. Because of the monotonous nature and long duration of each experiment (approximately 24 hr), subjects were allowed (and encouraged) to take breaks whenever they so desired. All the subjects completed their respective experiments in 14–16 sessions over a span of about 2 weeks.

**Procedure**

In each of the four experimental conditions, the task was a 2AFC letter-identification task in which the subject had to indicate whether the letter at the target position was “o” or “x”, irrespective of the flankers if they were present.

Each experimental condition consisted of 10 blocks with 1,100 trials per block. Thus, the entire experiment consisted of 40 blocks with the four experimental conditions randomly intermixed. For the two unflanked conditions, a white-on-black letter was presented on each trial in a Gaussian luminance noise field. For the two flanked conditions, three white-on-black letters (target + one flanker on each side) were presented in the noise field. The target was always at the center of the noise field. The noisy stimuli were presented at a viewing distance of 154 cm for the foveal viewing conditions and at 105 cm for the peripheral viewing conditions. In the peripheral viewing condition, subjects fixated at an LED 10° above the center of the noise field. The first 100 trials in each block were calibration trials in which the letter contrast was dynamically adjusted using the QUEST procedure (Watson & Pelli, 1983) as implemented in the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) to estimate a “calibrated” threshold contrast for reaching an accuracy level of 75%. The remaining 1,000 trials within a block were divided into 5 subblocks of 200 trials each, and QUEST was reinitialized to the calibrated value at the beginning of each subblock. During the initial 100 calibration trials, the standard deviation of the prior distribution of the threshold value was set to 5 log units (practically a flat prior), but for each subblock, the prior was narrowed to a standard deviation of 1 log unit. This restricted the variability of the test contrast but still allowed adequate flexibility for the procedure to adapt to the observer’s continuously improving threshold levels.

Prior to the main experiment, letter acuity was measured separately for the foveal and peripheral viewing conditions. The subjects were asked to identify any of the 26 lowercase letters, while the size of the presented letter was varied using QUEST to achieve an identification accuracy of 79%.

**Stimuli**

The stimulus for each trial consisted of either a white-on-black letter (unflanked) or three white-on-black letters (flanked) added to a Gaussian, spectrally white noise field of 128 × 128 pixels. Before being presented to the observers, each pixel of this noisy stimulus was enlarged by a factor of 2, such that four screen pixels were used to render a single pixel in the stimulus. This was done to increase the spectral density of the noise. The noise contrast was fixed at 25% rms. In the foveal viewing conditions (distance = 154 cm), the noise field was of size 3.2° with a spectral density of 39.8 μdeg⁻², whereas in the peripheral viewing conditions (distance = 105 cm), the noise field was of size 4.7° with a spectral density of 85.5 μdeg⁻². The mean luminance of the noisy background was 19.8 cd/m². In the flanked condition, the target and flankers had the same contrast.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Fovea letter size</th>
<th>Periphery letter size</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.B.</td>
<td>0.15°</td>
<td>0.81°</td>
</tr>
<tr>
<td>L.M.</td>
<td>0.20°</td>
<td>1.14°</td>
</tr>
<tr>
<td>O.R.</td>
<td>0.18°</td>
<td>0.89°</td>
</tr>
</tbody>
</table>

Table 1. Stimuli letter size (x-height) for each subject in units of degrees of visual angle.
The letter stimuli were presented in “Arial” font (Mac OS 9). The target letters were lowercase “o” and “x”. The flanking letters were randomly chosen from the following set: “a”, “c”, “e”, “n”, “r”, “s”, “u”, “v”, and “z”. The flanking letters did not include the target letters and were chosen such that none of them had any ascenders (e.g., “b”) or descenders (“p”) or were of unusual width (“w”, “i”). Letter size was set to twice the subject’s letter acuity at the respective eccentricity. The letter size (the height of a lowercase “x”) used for each of our subjects is shown in

Figure 4. (A) Examples of stimuli used in the four experimental conditions. (B) Timing of stimuli presentation: (1) fixation beep immediately followed by a fixation screen for 500 ms, (2) stimulus presentation for 250 ms, (3) subject response period (variable) with positive feedback beep for correct trials, and (4) 500 ms delay before onset of next trial.
Table 1. The center-to-center letter spacing in the flanked conditions was 1 x-height. The contrast of the target letter (and the flankers, if any) was adjusted with a QUEST procedure as described in the Procedure section. Figure 4A depicts examples of the noisy stimuli used in the four experimental conditions.

The stimuli were displayed in the center of a 19-in. CRT monitor (Sony Trinitron CPD-G400), and the monitor was placed at a distance of 105 or 154 cm (depending on the experimental condition) from the subject. After calibration and gamma-linearization, the monitor has 11 bits (2,048 levels) of linearly spaced contrast level. The experiments and the monitor were controlled from a Mac G4 running OS 9.2.2. All 11 bits of the contrast levels were addressable to render the noisy stimulus for each trial. This was achieved by using a passive video attenuator (Pelli & Zhang, 1991) and a custom-built contrast calibration and control software implemented in MATLAB. Only the green channel of the monitor was used to present the stimuli.

The stimuli were presented according to the following temporal design: (1) a fixation beep immediately followed by a fixation screen for 500 ms, (2) stimulus presentation for 250 ms, (3) subject response period (variable) with positive feedback beep for correct trials, and (4) 500 ms delay before onset of the next trial (see Figure 4B).

At the end of each trial, the following data were collected for subsequent data analysis: the state of the pseudorandom number generator used to produce the noise field, the identity and contrast of the target letter, the identity of the flankers (only for the two flanked conditions), and the response of the subject.

Results and discussions

Threshold SNR

Figure 5A depicts the log SNR threshold (contrast energy vs. noise spectral density) or \( \log(EN/N) \) for the four experimental conditions. We use SNR thresholds instead of the contrast thresholds to allow direct comparison between the two viewing conditions, which had different stimuli sizes and noise spectral densities. For all subjects, there was a significant threshold elevation in the periphery-flanked condition as compared with the periphery-unflanked condition, showing that we had indeed induced crowding in the periphery. In sharp contrast, there was a small but significant opposite effect in the fovea: The flankers actually aid recognition. As it will transpire from the results of the feature utilization zone (Figures 7 and 8), the facilitation appears to be the result of a reduction in intrinsic spatial uncertainty. The presence of flankers may act as guideposts marking the target position.

The human \( EN/N \) threshold can also be compared to the ideal-observer model described earlier in the Feature maps section. We will use this ideal-observer model to evaluate the human feature maps. Recall that the ideal-observer model optimally performs the letter-identification task without flankers but with a spatial uncertainty equated to \( \pm 1.5 \) x-height in both horizontal and vertical dimensions. Table 2 summarizes the total efficiency (defined as the SNR threshold \( [EN/N] \) of the ideal-observer model divided by the SNR threshold of the human observer) for the four experimental conditions. These values are higher than what would be expected from the literature (e.g., Nasanen & O’Leary, 1998; Pelli et al., 2004; Solomon & Pelli, 1994; Tjan, Braje, Legge, & Kersten, 1995) because our ideal-observer model is limited by spatial uncertainty, thus making it suboptimal with respect to the actual stimuli, which did not have any spatial uncertainty.

First-order classification images

The reconstructed classification images for subjects BB, LM, and OR are depicted in Figure 5B. The images were reconstructed separately for the four experimental conditions; the subimages for the individual stimulus–response categories were not combined as we have previously argued (Tjan & Nandy, 2006), and we limited our analyses to the error-trial subimages. We found that, with the exception of the periphery-flanked condition for subject OR, the signal-clamped classification-image technique was able to reveal a first-order template associated with the presented signal for all conditions. The templates were virtually identical for the two foveal conditions—we can clearly see the “ring” for the “o” and the “dot” for the “x”. The “dot” corresponds to the region where the two oblique strokes of the “x” intersect. In the periphery-unflanked condition for all subjects, the templates were of a lower contrast than the fovea-unflanked condition (see Table 3, first row vs. second row) but were nevertheless undistorted. They also had the same structural characteristics as the foveal templates. Computational simulations in Tjan and Nandy (2006) showed that if an observer mechanism uses a template that differs from the presented stimulus, it is the perceptual template, not the presented stimulus, which forms the negatively correlated component in the classification image of the error trials. Hence, we conclude that there was no uncalibrated sampling disarray (contrary to Hess & Field, 1993; Hess & McCarthy, 1994) at the scale of the letters in the periphery.

More important, for subjects BB and LM in the crowding condition (periphery flanked), the template for the letter “o” was undistorted as well, albeit of a lower contrast (combined rSNR for “o” and “x”: 26.3 [BB], 35.6 [LM]). This finding suggested that perceptual templates in the periphery, even under crowding, are not distorted and that crowding could not be attributed to aberrations in the template.

The contrast of the classification images in the periphery-flanked condition was substantially lower than that in the
Figure 5. (A) Log thresholds in SNR (contrast energy $E$ divided by noise spectral density $N$) for the four experimental conditions. There is a significant increase in $E/N$ in the periphery-flanked condition as compared with the periphery-unflanked condition. There is also a small but significant opposite effect in the fovea. Note that the standard error bars are smaller than the plot symbol. (B) First-order classification images. The raw images have been filtered by a Gaussian kernel (space constant = 1.4 pixels) to aid visualization. The numbers indicate the number of trials for the corresponding stimulus–response category. For the fovea conditions with small letter size, only the central portion of the classification images is magnified and shown.
The mean of the old list with the 10 largest items removed will be slightly lowered but may not reach statistical significance according to a \( z \) test. For this reason, we only analyze the error trials.

In contrast to results in the periphery, the errors in the fovea were entirely driven by the noise in the target region. There was no effect of the flankers. Neither the \( z \)-score nor the \( t \)-test maps show any significant correlation between pixels in the flanker locations and the error response. Most of the pixels with significant correlations were located in the target position (i.e., congruent with the classification images).

### Feature utilization zones

Excluding feature mislocalization, are there other causes of crowding? In this subsection as well as in the next, we will analyze the second-order correlation structures in the residual noise fields. A residual noise field (Equation 5) is a noise field that does not contain the flankers and is orthogonal to the corresponding first-order classification subimage. As mentioned earlier, the method of signal-clamped classification image separates the mechanism selective to the presented target from the mechanism selective to the alternative target by their relative effects on the first-order classification subimages. This is because of the fact that a high-contrast signal reduces the intrinsic spatial uncertainty of only the mechanism selective to the presented signal. With high intrinsic spatial uncertainty in peripheral vision, the resulting first-order signal-clamped classification sub-images from error trials will be dominated by the perceptual template (spatiotemporal average of features) of the mechanism selective to the presented target. By projecting out this first-order classification subimage (Equation 5), the residual noise fields will contain the higher order correlation structures associated with the mechanism selective to the target that was not presented and, thus, not spatially clamped by the stimulus. The spatial dispersion of these correlation structures from the target position provides a measure of the intrinsic uncertainty and the spatial range within which features were detected and utilized to affect response. This spatial range is referred to as a feature utilization zone and is defined by Equation 14. It is important to bear in mind that this feature utilization is not to be confused with feature mislocalization estimated in the pre-

### Table 3. rSNR of the classification images

<table>
<thead>
<tr>
<th></th>
<th>B.B.</th>
<th>L.M.</th>
<th>O.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fovea unflanked</td>
<td>175.4</td>
<td>261.2</td>
<td>108.8</td>
</tr>
<tr>
<td>Periphery unflanked</td>
<td>150.9</td>
<td>214.1</td>
<td>82.1</td>
</tr>
<tr>
<td>Fovea flanked</td>
<td>152.5</td>
<td>154.4</td>
<td>78.9</td>
</tr>
<tr>
<td>Periphery flanked</td>
<td>26.3</td>
<td>35.6</td>
<td>( \approx 0 )</td>
</tr>
</tbody>
</table>

The mean of the old list with the 10 largest items removed will be slightly lowered but may not reach statistical significance according to a \( z \) test. For this reason, we only analyze the error trials.

In contrast to results in the periphery, the errors in the fovea were entirely driven by the noise in the target region. There was no effect of the flankers. Neither the \( z \)-score nor the \( t \)-test maps show any significant correlation between pixels in the flanker locations and the error response. Most of the pixels with significant correlations were located in the target position (i.e., congruent with the classification images).
Figure 6. Flanker analysis results: (A) z-score maps (Z_{OX}, upper right box; Z_{XO}, lower left box) thresholded at $\alpha = .05$ ($|Z| > 1.96$) indicate the presence (positive contrast) or absence (negative contrast) of significant features that bias an observer’s response; (B) t-test maps, also thresholded at $\alpha = .05$, directly compare the error trials of the two response categories. The “hot” regions indicate features that bias toward “o,” whereas the “cold” regions are indicative of feature that bias toward “x.” For the fovea conditions, only the central portion of the classification images is magnified and shown.
Figure 7. Estimated ROI\textsubscript{opt} or feature utilization zones. The significance (mean of log $p$ values) of the different candidate ROIs are color coded, and these color-coded regions are superimposed in ascending order of significance; the most significant regions (demarcated by the blue dotted lines) represent the optimal ROIs. The positions of the target and the flankers are superimposed to give a sense of the extent of the utilization zones.
vious section. The two are independent because the residual noise fields do not contain the flankers.

The feature utilization zones are depicted in Figure 7. The significance (mean log $p$ values, Equation 13) of the different candidate ROIs (Equation 12) are color coded, and these color-coded regions are superimposed in ascending order of significance (descending order of mean log $p$ values); the most significant region (demarcated by a blue dotted line) represents the optimal ROI (Equation 14) and, hence, the feature utilization zone. The positions of the target and the flankers are superimposed on the maps to give a sense of the extent of the utilization zones.

Figure 8 summarizes the horizontal extent of the feature utilization zones and the significance level of the optimal ROIs for the different experimental conditions for each of our subjects. The significance level (mean log $p$ values) associated with the optimal ROI can be taken to indicate the sensitivity and trial-to-trial consistency of feature detection. By this measure, there is no coherent pattern between subjects in terms of feature detection (Figure 8A). Comparing the two periphery conditions, we found that feature detection is numerically weaker in the flanked condition for subject BB and that there is no difference for subjects LM and OR. This result is consistent with most theories of crowding—that the deficiency lies in second-stage feature integration and not in the detection of simple features (e.g., He et al., 1996; Pelli et al., 2004).

For subject BB, the feature utilization zone (Figure 7, demarcated in blue) shrinks vertically in the flanked conditions for both foveal and peripheral presentations. Thus, for this subject, who was our most experienced subject and who had performed in similar experiments for over 120,000 trials, the zone of feature utilization is quite well bounded around the target region even under crowding. In both the fovea and the periphery conditions, the presence of flankers led to a sizable reduction in the subject’s intrinsic spatial uncertainty. For subject OR, the utilization zone similarly shrinks vertically in the fovea-flanked con-
dition. In the periphery, the zone of utilization in the flanked condition remains roughly the same as in the unflanked condition. For subject LM, the utilization zones remain roughly the same under flanking in both fovea and periphery conditions. In general, the utilization zone either shrunk vertically, orthogonally to the letter arrangement, or remained unchanged in the flanked condition as compared with the unflanked conditions at the corresponding eccentricity. Relative to the stimulus size, the utilization zones in the periphery-flanked (crowded) condition were not much larger than those in the fovea-flanked (noncrowded) condition. As such, crowding cannot be attributed to an increase in the intrinsic spatial uncertainty associated with feature utilization beyond any feature mislocalization induced by the flankers. This conclusion is further strengthened by a lack of change in the horizontal extent of the feature utilization zones between the crowded and noncrowded conditions in the periphery (Figure 8B).

The relatively larger horizontal extent of the feature utilization zones in the fovea compared to the periphery (subjects BB and OR) in units relative to the stimulus size deserves further commentary. The feature utilization zone reveals the image locations from which simple features were extracted, excluding any direct influence of the flankers. It does not, however, tell us what happens after those features have been extracted. In the fovea, although the spatial uncertainty of a letter stimulus may be (hence, a large feature utilization zone), the relative uncertainty between the letter features can still be very low. For example, a widely separated pair of “/” and “\” may not be put together to form “X” in the fovea. In contrast, despite having a smaller feature utilization zone, the relative spatial uncertainty between features may still be very large in the periphery. It is also possible that feature utilization in the periphery is sparse, such that the mere presence of a simple feature may be sufficient to trigger a false alarm. A zone with either sparse feature utilization (improvised feature integration) or high relative spatial uncertainty between features (improper feature integration) cannot effectively exclude the influence from flankers.

**Second-order feature maps**

Using the feature utilization zones (optimal ROI) identified in the previous section, we calculated the second-order (pairwise) correlations between pixels in the residual noise fields (see the Feature maps section). Because the masking noise was uncorrelated white noise, any significant correlations detected in the residual noise field can only be due to the coincidental features in the noise that were consistently detected by the observer as features that affected the behavioral response. Unlike the first-order classification images, which reveal the spatial configuration of features averaged across trials but not the features themselves, these correlation maps reveal the second-order structure of the extracted subtemplate features. By “feature”, we specifically mean a fragment of the target (a group of pixels) that is detected as a whole within a trial (as opposed to pixels averaged across trials).

The second-order feature maps for subjects BB, LM, and OR are shown in Figure 10. For purposes of comparison, the feature maps calculated from an ideal-observer model performing the same 2AFC letter-identification task in the unflanked conditions are shown in Figure 9. The ideal-observer model (Tjan & Nandy, 2006) used the letter stimuli corresponding to the ones used in the human experiment and is limited by a considerable amount of spatial uncertainty (±1.5 x-height). The positive and negative correlation zones from the ideal-observer maps at \( rZ = ±1.0 \) are overlaid on the human correlation maps as black (+) and white (−) contours, respectively. One important fact to bear in mind is that the ideal-observer model uses the entire letter as a single feature for its pattern matching operations. If a subject were using a fixed set of letter fragments as features (e.g., a pair of parallel bars to detect the letter “o”), then the positive part of the feature map of the subject would be strictly a subset of the significant positive correlations found in the corresponding ideal-observer feature map. It is also possible that a subject used features that are considered inefficient by the ideal-observer model. In general, we may not expect any good match between the human and ideal-observer feature maps, but a comparison between the two is informative.

Let us first consider the maps for subject BB. In both the fovea conditions regardless of flanking, the subject (Figure 10: rows 1 and 2, columns 1 and 2) seems to be using a very similar set of second-order correlations as that of the ideal-observer model, suggesting that the subject used the entire letter as a unitary template. The resemblance is more striking for the letter “x” than for the letter “o”. In the periphery, however, the subject seems to be using a different feature set as compared with the ideal-observer model. The feature set in the periphery-unflanked condition (Figure 10: row 1, columns 3 and 4) is sparser, but it still appears to draw from the more complete set used by the ideal-observer model. For example, whereas the ideal-observer model utilizes a complete set of diagonal correlation for the letter “x”, the subject seems to be using correlated fragments at the extremities of a diagonal. In the periphery-flanked condition (Figure 10: row 2, columns 3 and 4), the deviation from the ideal-observer model is more pronounced.

To quantify feature detection and utilization for all of our subjects, we compare a human feature map with the corresponding ideal-observer feature map in terms of the three quantities defined in the Feature maps section: quality of match (\( Q_m \)) between the human and ideal-observer feature maps, fraction of the ideal-observer features that are also used by the human observer (\( U \), or feature utilization), and the fraction of the human features that are also
features for the ideal-observer model ($V$, or feature validity). Recall that for $U$ and $V$, we consider only positively correlated features with $rZ > +1.0$. Recall also that we will separately analyze the mechanisms used for detecting “o” from those used for detecting “x”. The results are summarized in Figure 11.

**Fovea conditions**

The quality of match ($Q_m$) will be close to 0.5 if a human feature map bears no resemblance to the ideal-observer feature map. In the fovea, flanking the target led to better performance in terms of contrast threshold; it also led to a statistically significant increase in $Q_m$ for all
Figure 10. Second-order feature maps obtained from the human observer’s data overlaid with positive (demarcated by the black contour lines) and negative correlation zones (white contour lines) from the corresponding ideal-observer feature maps (Figure 9). The axes of the maps are in units of x-height. $|rZ| > 1.96$ corresponds to an $\alpha$ level of .05.
mechanisms in all subjects. In the fovea-flanked condition, all $Q_m$ values are significantly above 0.5, with the two lowest $Q_m$ values of 0.64 and 0.67 coming from the “o” mechanisms of subject BB and OR, respectively. Visually inspecting the feature maps in Figure 10 confirms this quantification.

With the exception of the “x” mechanism for BB, feature utilization ($U$) for all subjects was below 20%, with an average of about 10% (excluding BB [“x”]). This is a clear piece of evidence that human observers do not use a whole letter as a single unitary feature. This result is comparable to the total efficiency measured with respect to the ideal-observer model (see Table 2). Flankers had a mixed impact on feature utilization for subject LM, increasing $U$ for the “x” mechanism and decreasing it for the “o” mechanism. For both BB and OR, flanking the target led to an overall increase of feature utilization. The effect of flankers on feature utilization being small and somewhat equivocal is consistent with the fact that the effect of flankers on threshold was small in the fovea conditions.

With one exception, feature validity ($V$) for both mechanisms of all subjects was less than 50% (average $V = 28\%$). This means that more than half (and in some cases, substantially more than half) of the features used by a human observer are not features considered by an ideal-observer model. Inspecting Figure 9 suggested that many of these spurious features were not as egregious as the low values of $V$ might suggest. The human features are often slightly displaced versions of the ideal-observer features. Take as an example the feature map of the “o” mechanism for BB in the fovea-unflanked condition. The locations of the positive correlations seem to shift toward the center of the map as compared with the ideal-observer overlay (Figure 10, first row, first column). The human feature set in this case was consistent with an “o” that is narrower by about 10–20% than the template used by the ideal-observer model. For the same subject, feature validity for the “o” mechanism improved significantly from 22% to 33% in the flanked condition.

**Periphery conditions**

In the periphery-flanked condition, the quality of match ($Q_m$) between human and ideal-observer feature maps was at chance (0.5) for subject OR and close to chance for subject LM. Remarkably, the quality of match for BB remained high for both flanked and unflanked conditions. Flanking the target in the periphery, which led to a large increase in contrast threshold, either reduced $Q_m$ (for OR, who had a SNR threshold elevation of 1.5 log units) or had no effect on $Q_m$ (for subjects BB and LM, for whom SNR threshold elevation was about 0.6 log units).
The negative effects of flanking were more consistently observed in terms of feature utilization \((U)\). For all subjects, flanking led to a significant reduction in \(U\) and a corresponding decrease in feature validity \((V)\). In general, for both the fovea and periphery conditions, the effect of flanking on feature validity \((V)\) mirrored its effect on feature utilization \((U)\). That is, a decrease in \(U\) was almost always paired with a corresponding decrease in \(V\). In the periphery conditions, this means that flankers reduced letter-identification performance not only by reducing the number of valid features used by a human subject but also by increasing the number of invalid features.

**General discussion**

We developed a series of classification-image techniques to investigate the nature of crowding without presupposing any model of crowding. Our findings can be summarized as follows: (1) crowding significantly reduced the contrast of first-order classification images, although it did not alter the shape of the classification images; (2) errors during crowding were strongly correlated with the spatial structures of the flankers that resembled those of the erroneously perceived targets; (3) crowding did not have any systematic effect on intrinsic spatial uncertainty, nor did it suppress feature detection; and (4) analysis of the second-order statistics of the classification images revealed that crowding reduced the amount of valid features used by the visual system and, at the same time, increased the amount of invalid features used.

Our data are informative about the three proximal causes of crowding concerning interactions between features from the target and the flankers, which we have referred to as masking, inappropriate feature integration, and source confusion (see the **Introduction** section). Our findings, however, do not directly address whether a lack of spatial resolution in the attention mechanism is the root cause of visual crowding (He et al., 1996; Intriligator & Cavanagh, 2001; Leat et al., 1999; Strasburger et al., 1991; Tripathy & Cavanagh, 2002).

**Source confusion**

Our data clearly support the view that source confusion between target and flankers is a main source of crowding (Krumhansl & Thomas, 1977; Strasburger, 2005; Strasburger et al., 1991; Wolford, 1975). In the periphery, there was strong correlation between a subject’s erroneous responses and parts of the flankers that resemble target features (Kooi, Toet, Tripathy, & Levi, 1994). Such correlation was completely absent in the fovea when acuity-scaled letter targets were similarly flanked.

We used the spatial distribution of second-order correlations present in the noise fields to measure the spatial region within which features are extracted by the visual system. The feature utilization zone thus determined was a graded property. Two findings concerning the feature utilization are particularly interesting. First, crowding did not systematically affect the horizontal extent of the feature utilization zone (see **Figure 8B**). Second, relative to target size, the feature utilization zone appeared to be larger in the fovea than in the periphery. Combined with the results of the flanker analysis, a consistent view of feature processing in the periphery emerged. In our experiments, flankers had the same contrast as the target. The fact that high-contrast features from the flankers were confused with those from the target cannot be attributed to the periphery being more promiscuous about where to look for the relevant features. Both foveal and peripheral mechanisms look for features outside the region of the target. Both are equally likely to be “tricked” by the low-contrast spurious features in the noise field. However, the fovea can distinguish between a high-contrast feature from the flankers from one that is extracted from the target, whereas the periphery cannot. In other words, both foveal and peripheral mechanisms know where to “look” for features, but once this is accomplished, the peripheral mechanism no longer knows where a feature came from.

If by “spatial attention” one means the control in defining the spatial region from where to extract features, then our data are inconsistent with the view that crowding is due to a limited spatial resolution of the attention system. On the other hand, if by “spatial attention” one means the ability and precision to maintain the location tag of a detected feature, then our data are consistent with the attention hypothesis.

**Masking**

Consistent with other studies, our data do not support masking being the cause of crowding (Chung et al., 2001; Levi et al., 2002; Majaj, Pelli, Kurshan, & Palomares, 2002; Pelli et al., 2004). Specifically, we found no evidence that crowding suppressed feature detection. If fewer features were being detected, then there would be fewer consistent second-order correlations in the error-trial noise fields, which was not the case (Figures 8B and 10). Consistent with this result is our finding that although crowding led to a decrease in the utilization of valid features, it also led to an increase in the proportion of invalid features (Figure 11). A masking account will predict that the set of invalid features are random from trial to trial and, as such, will not be detected by our second-order correlation analysis and be classified as invalid features. Therefore, we should see a reduction in \(U\) but no change in \(V\) if masking was at work.

Our results, however, do not reject the idea that suppression of feature detection ever occurs along the visual-processing hierarchy. For example, it is possible that surround suppression leads to a degradation of feature detection in the early visual stages. To compensate, the later stages try
to infer the missing features by an error-prone process. For this scenario to be consistent with our data, the error-prone process of feature inference must be relatively deterministic with respect to the stimuli; otherwise, it will not be possible to generate any consistent second-order correlations associated with the invalid features, as we found in our data.

Inappropriate feature integration

Our method of analysis identifies features in terms of their second-order statistics. A feature, in the strict sense of our analysis, refers to a pair of correlated pixels, which we can think of as the most basic form of a feature. A full analysis of feature integration will require analysis beyond the second order. Nevertheless, a partial analysis is possible with our current data set. Errorneous feature integration by itself will lead to an increase in the amount of invalid second-order features. We found a decrease in feature validity during crowding, which is consistent with this prediction (Figure 11, periphery, third row). However, we also found a decrease in the amount of valid features (Figure 11, periphery, second row), which is not predicted by spurious feature integration per se. To account for our result, the process of inappropriate feature integration (Levi et al., 2002; Pelli et al., 2004) must also somehow suppress the detection of valid features. This can be the case if the process of integration is a competitive one, a scenario that is highly probable. For instance, the idea of association fields (Field, Hayes, & Hess, 1993) for contour completion is an example of a competitive feature-integration process; there are situations when the visual system can complete a contour one way or another, but never both, although the decision is ambiguous at a local detector level. A phenomenon known as bias competition found in V4 and higher cortical areas (Chelazzi, Miller, Duncan, & Desimone, 2001; Desimone & Duncan, 1995; Luck, Chelazzi, Hillyard, & Desimone, 1997; Moran & Desimone, 1985; Reynolds, Chelazzi, & Desimone, 1999), where disjointed patterns in the receptive field of a single neuron will “compete” for the control of the neuron’s firing rate, may serve as a neural substrate for the competitive feature-integration process. The ubiquitous divisive normalization in the visual cortex (e.g., Carandini, Heeger, & Movshon, 1997; Heeger, 1992; Legge & Foley, 1980) provides a computational basis of a competitive process.

Beyond a two-letter discrimination task

We believe that by deliberately introducing a large amount of spatial uncertainty to the stimulus, our method of signal-clamped classification images and the accompanying analysis procedures to infer the perceptual template and the subtemplate features (in terms of their second-order statistics) could be suitably and practically extended to nAFC tasks by using $n \times N$ trials where $2N$ is the number of trials needed to form a classification image of sufficient quality in a 2AFC task ($N \approx 6,000$ for the “o” vs. “x” task according to the analysis in Tjan & Nandy, 2006). In contrast, such a task would require $n^2 \times N$ trials under the standard classification-image paradigm. This enormous saving in the number of trials results from the fundamental premise of the signal-clamped method—that under high spatial uncertainty, only the mechanism selective to the presented target can form a clear first-order classification image in the error trials. Contributions from all the other mechanisms that “false-alarmed” in the error trials are spatially dispersed due to the high spatial uncertainty in the stimulus (and/or spatial uncertainty intrinsic to the observer). Hence, in a 26-letter identification task, a first-order characterization of the mechanism for detecting the letter “A” among other letters can be obtained by averaging the noise fields from all the error trials when “A” was presented (i.e., when the observer committed a miss on the letter “A”) regardless of the observer’s response (as long as it is not “A”). This will yield a clear first-order template for “A” with nonsignificant contributions from the other mechanisms. To obtain a second-order characterization of “A”, we would collect all the error trials when the observer false-alarmed on “A” (responded “A” when another letter was presented) and apply the second-order feature-map and optimal ROI analyses on the corrected noise fields (i.e., after projecting out the average of this collection of noise fields). For a 26-letter identification task, we realize that $26 \times 6,000$ is still a very large number of trials, but it is certainly far better than $26^2 \times 6,000$. Finally, to fully investigate letter crowding, testing 10 letters will be more than sufficient. It is also not necessary to exclude target letters from the flanker positions. To simplify the analysis, all that is required is that the target letter is different from the flanking letters in a trial.

Summary

With respect to the three proximal causes of visual crowding that have been hypothesized, our data strongly support the source-confusion or feature-mislocalization hypothesis and, at the same time, argue against a front-end version of the spatial attention account. Our data also support the inappropriate feature-integration hypothesis but require feature integration to be a competitive process. Our data reject the feature-masking hypothesis; we do not rule out feature masking entirely but require that a suppressed feature-detection process be paired with a feature-inference process that leads to a consistent set of spurious features.

Acknowledgments

We thank the two anonymous reviewers for their very helpful comments. This research was supported in part by National Institutes of Health Grant EY016391 to BST.
Commercial relationships: none.
Corresponding author: Bosco S. Tjan.
Email: btjan@usc.edu.
Address: Department of Psychology, 3620 South McClinton
Parkway, Los Angeles, CA 90089-1061, USA.

Footnotes

1 Pelli et al. (2004) distinguished the study of Treisman and Schmidt (1982), which coined the phrase “illusory conjunction,” from most of the other studies of this phenomenon. Pelli et al. argued that the experimental conditions used in the study of Treisman and Schmidt led to temporal crowding, whereas most of the later studies used conditions that led to ordinary (spatial) crowding.

2 We make the distinction between an ideal observer and an ideal-observer model. Whereas an ideal observer is defined solely with respect to a given task and its stimuli, an ideal-observer model, in contrast, includes also the assumed limitations of a human observer. That is, an ideal observer is optimal with respect to a task and its stimuli; an ideal-observer model, on the other hand, is optimal with respect to the task, stimuli, and the explicitly assumed limitations of the human observer.

References


areas (18 and 19) of the cat. Journal of Neurophysiology, 28, 229–289. [PubMed]


