Crowding and attention in a framework of neural network model

Endel Põder

Institute of Psychology, University of Tartu, Tartu, Estonia

In this article, I present a framework that would accommodate the classic ideas of visual information processing together with more recent computational approaches. I used the current knowledge about visual crowding, capacity limitations, attention, and saliency to place these phenomena within a standard neural network model. I suggest some revisions to traditional mechanisms of attention and feature integration that are required to fit better into this framework. The results allow us to explain some apparent theoretical controversies in vision research, suggesting a rationale for the limited spatial extent of crowding, a role of saliency in crowding experiments, and several amendments to the feature integration theory. The scheme can be elaborated or modified by future research.

Introduction

Visual crowding is a deterioration of the perception of a target object caused by other objects nearby (Bouma, 1970; Andriessen & Bouma, 1976). Crowding is primarily observed in the visual periphery where it is a main factor that limits pattern recognition (Pelli & Tillmann, 2008; Levi, 2008).

There are many theoretical ideas about the mechanisms of crowding and not much agreement among the researchers (Balas, Nakano, & Rosenholtz, 2009; Greenwood, Bex, & Dakin, 2009; Van den Berg, Roerdink, & Cornelissen, 2010; Herzog, Sayim, Chicherov, & Manassi, 2015; Francis, Manassi, & Herzog, 2017; Harrison & Bex, 2017; Manassi & Whitney, 2018; Rosenholtz, Yu, & Keshvari, 2019). A long-lasting question is about the role of attention (Intriligator & Cavanagh, 2001; Pelli, Palomares, & Majaj, 2004; Põder, 2006; Strasburger, 2007). Whereas some authors see crowding as an implication of certain properties of attention, others reject any important relationship between these.

In this article, I attempt to clarify relations between visual crowding and different forms of attention. I analyze crowding and attention in the context of a simple computational framework for vision—a feedforward convolutional neural network (CNN). It appears that several puzzling results from crowding experiments can be naturally explained by simple attentional mechanisms combined with standard feedforward network.

A classic neural network model

Based on much neurobiological data and state-of-the-art machine vision, we can describe a standard object recognition system as a hierarchical feedforward neural network (Hubel & Wiesel, 1965; Fukushima, 1980; Riesenhuber & Poggio, 1999; Krizhevsky, Sutskever, & Hinton, 2012; Kubilius, Bracci, & de Beeck, 2016; Figure 1). It consists of a number of feature maps at several levels of processing. A feature map is an array of local "feature detectors." Features at higher levels become more complex in terms of local pixel values but more relevant for object recognition tasks. A number of different features is larger at higher levels. Some extent of spatial pooling occurs at every level—receptive fields become larger, and precise absolute position of features is discarded. A few of the highest levels consist of nonspatial feature maps. Crowding is a natural consequence of spatial pooling within this architecture.

A puzzle of Bouma's law

Many studies of visual crowding have found that crowding zones have radius approximately 0.5 of the eccentricity of the target object (Bouma, 1970; Pelli et al., 2004). Assuming that crowding is an effect of spatial pooling, this value should reveal the size of respective receptive fields. On the basis of these findings, crowding has been frequently attributed to a single level of processing where the size of receptive fields best fits the given dependence on eccentricity (V2, or V4; Freeman & Simoncelli, 2011; Motter, 2018). However, this simple account of crowding seems to be at odds with some knowledge about object recognition systems.

Position-invariant object recognition with feedforward neural networks apparently requires

Citation: Põder, E. (2020). Crowding and attention in a framework of neural network model. *Journal of Vision*, *20*(13):19, 1–10, https://doi.org/10.1167/jov.20.13.19.

https://doi.org/10.1167/jov.20.13.19 Received April 23, 2020; published December 29, 2020 ISSN 1534-7362 Copyright 2020 The Authors

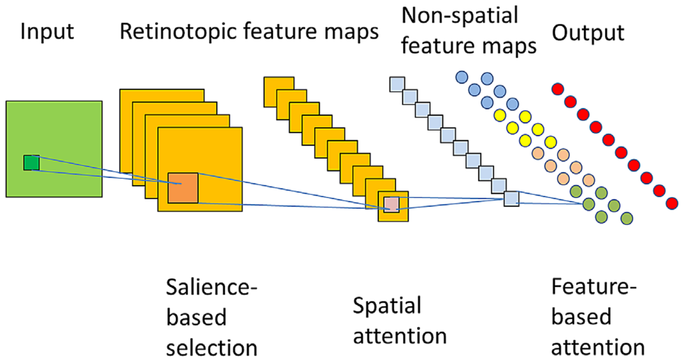

Figure 1. A standard neural network model of vision. Approximate levels of basic attention mechanisms are indicated.

cumulative spatial pooling over the full visual field. Standard CNNs contain several steps of pooling, with the final output collapsed across spatial coordinates. This implies some crowding-like interference from irrelevant objects located anywhere in the visual field. In reality, available data indicate that zones of interference in CNNs may extend over the whole input image, and target-distractor distance has only a modest effect (Lonnqvist, Clarke, & Chakravarthi, 2020). Similarly, a systematic increase throughout the levels and very large receptive fields at the highest levels have been found in the primate visual system (Gross, Rocha-Miranda, & Bender, 1972; Rolls, 2000). Given these observations, it seems puzzling that spatial extent of crowding in human vision is limited to a relatively small fraction of the visual field with radius 0.5*E*.

An explanation quite naturally follows from Intriligator and Cavanagh's (2001) idea that crowding reflects a maximum resolution of spatial attention. When a "spotlight" of attention has radius of 0.5*E*, we can voluntarily select available information within that zone and exclude everything outside of it (Figure 2). Therefore attentional selection can eliminate any potential crowding caused by stimuli at larger distances. This role of spatial attention has been demonstrated in the inferotemporal (IT) cortex of monkeys (Zhang, Meyers, Bichot, Serre, Poggio, & Desimone, 2011). Because the "spotlight" of attention cannot be reduced beyond 0.5*E*, crowding within that zone cannot be eliminated.

Still, there exist some conditions when crowding zones may be either much smaller (e.g., Põder, 2006), or much larger (Vickery, Shim, Chakravarthi, Jiang, & Luedeman, 2009). "Too large" crowding zones can be likely explained by incomplete focusing of attention in certain conditions. Reduced extent of crowding is the topic of the next part.

The idea to relate the limited extent of crowding to attention leaves unexplained why the minimum radius of the attentional spotlight is 0.5*E*. Possibly, that could be explained by some properties of natural environment or limits of brain capacity.

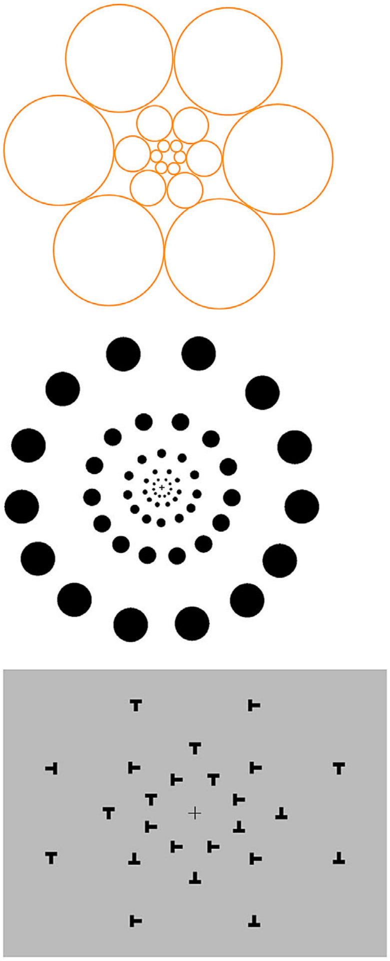

Figure 2. Crowding zones and spatial resolution of attention. Classic crowding zones with radius 0.5*E* according to Bouma's law (top), spatial resolution of attention (minimal spacing that allows attending individual items) according to Intriligator and Cavanagh (2001, center; reprinted with permission from Elsevier), "uncrowded" display of rotated Ts (all inter-object distances are at least 0.6*E*, bottom). While fixating the central cross, we can voluntarily attend every single letter and avoid crowding from others.

Finally, we must note that human peripheral vision and convolutional neural networks are not directly comparable. The standard CNNs with their built-in position invariance approximate object recognition in human central vision. In human peripheral vision, not only critical distance of crowding but also many other important characteristics vary with eccentricity (e.g., Strasburger, Rentschler, & Juttner, 2011). Up to now, very few studies have tried to implement these eccentricity dependencies in neural network models (Chen, Roig, Isik, Boix, & Poggio, 2017; Han, Roig, Geiger, & Poggio, 2020).

Saliency-based selection

It is well known that crowding is significantly reduced when a target differs from the flankers by some simple visual feature (for example, has a different color, e.g., Kooi, Toet, Tripathy, & Levi, 1994; Põder, 2007). This effect is even larger when the number of flankers is increased (Põder, 2006; Malania, Herzog, & Westheimer, 2007; Manassi, Sayim, & Herzog, 2013). With many homogeneous flankers, crowding virtually disappears. Põder (2006) proposed an explanation based on two mechanisms of spatial selection. In addition to voluntary attention with relatively low spatial resolution, there is a saliency-based bottom-up selection with much higher resolution. Most likely, it operates at lower levels of processing, possibly at V1 (Li, 1999; Li, 2002). Multiplication of the effects of these two selection mechanisms produces experimentally observed crowding zones (Figure 3).

However, this is not the full story. There is a top-down feature-based attention that can bias selection of salient objects. For example, it is not difficult to select either a blue or orange pop-out bar presented among homogeneous green bars (Figure 4). With a heterogeneous background, however, there is no pop-out, and selection of even a single target color is nearly impossible. Little is known of how this combining of bottom-up and top-down selection is implemented in neural networks.

Researchers have designed many different crowding displays that produce results not consistent with Bouma's law and simple pooling models (Livne & Sagi, 2007, 2010; Saarela, Sayim, Westheimer, & Herzog, 2009; Manassi, Sayim, & Herzog, 2013; Rosen & Pelli, 2015; Doerig, Bornet, Rosenholtz, Francis, Clarke, & Herzog, 2019; Figure 5). Frequently, a notion of grouping has been evoked as an explanation (Livne & Sagi, 2010; Herzog et al., 2015; Francis et al., 2017). However, grouping itself is not well understood, and its relation to crowding is ambiguous too. I suggest that saliency could be a more productive idea. There exist neurobiological data on saliency computation in biological vision (Knierim & van Essen, 1992; Sillito, Grieve, Jones, Cudeiro, & Davis, 1995; Nothdurft, Gallant, J. & Van Essen, 1999, Li, Their, & Wehrhahn, 2000), and several methods to compute saliency maps of images have been proposed (Itti & Koch, 2000; Zhang, Tong, Marks, Shan, & Cottrell, 2008; Bruce & Tsotsos, 2009). Relative saliency of the target and flankers could directly predict performance in crowding experiments.

It appears that that many "complex" results from crowding studies are naturally explained by a simple lateral-inhibition–based saliency model. Illustrative examples are given in Figure 6. Here, saliency of an item i was calculated as $s_i = e_i - \sum_j k_{ij} e_j$, where e_i is activation of an item before lateral inhibition

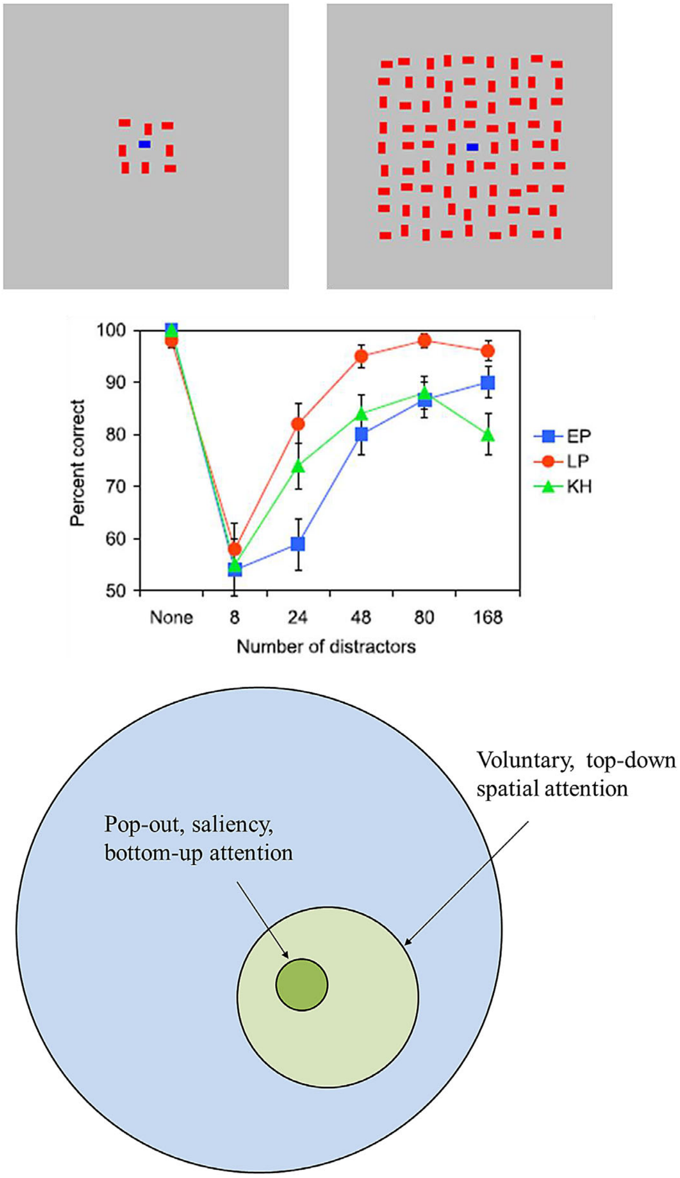

Figure 3. Reduced crowding with large number of homogeneous flankers (Põder, 2006). Examples of stimuli (top), experimental results (center), and illustrative model with two selection mechanisms (bottom).

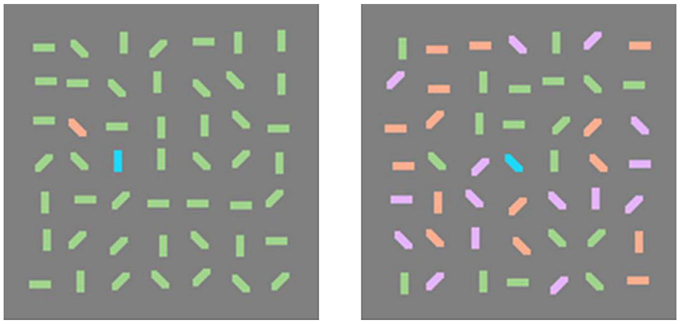

Figure 4. Combination of bottom-up saliency and top-down feature-based attention. When presented in the periphery, it is not difficult to attend to either orange or blue salient bar (left). With heterogeneous distractors, it is difficult to attend to a single blue bar (right).

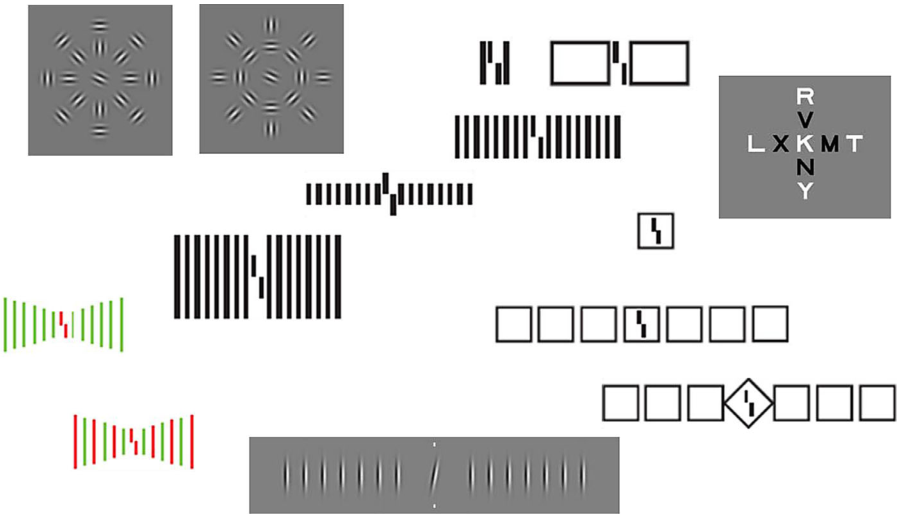

Figure 5. Various examples of stimuli from crowding experiments that appear to reject simple pooling models. Number, homogeneity, and regularity of flankers have strong effects on target perception (Malania et al., 2007; Saarela et al., 2009; Livne & Sagi, 2010; Manassi et al., 2013; Rosen & Pelli, 2015).

and k_{ij} is inhibitory interaction between items i and j.

To account for Põder (2006) results, we must assume that inhibitory interactions are stronger between items of the same color, as well as between those spatially close to each other. This mechanism suppresses signals from the flankers within a group of the same color, while differently colored target receives much less inhibition. With a larger number of flankers, the total suppression affecting each one is stronger. Suppression is less at the corners and edges of a group. Thus increasing the number of same-color flankers both reduces saliency of the flankers and moves salient flankers further from the target (Figure 6A). To reproduce the experimental results, the saliency model must be combined with an ordinary crowding zone/a window of top-down spatial attention (as shown in Figure 3).

Livne & Sagi (2007) found an effect of global configuration of flankers on crowding. For example, a smooth circle of flanking Gabors caused much less crowding than a sun-like configuration. Simple orientation- and proximity-based inhibition cannot explain these results. Interestingly, however, Cavanaugh, Bair, & Movshon (2002) have reported on an interaction of orientation and relative position in surround suppression in V1. In addition to the simple effect of orientation tuning, radially oriented surround patches caused stronger suppression compared to tangentially oriented ones. Applying this rule to Livne and Sagi (2007) stimuli produces a stronger inhibition for the target and less inhibition for the flanker signals in "sun" as compared to "smooth" configuration (Figure 6B).

Manassi et al. (2016) used two different shapes as flankers in a Vernier offset discrimination task. Using a row of flanking squares, the crowding effect decreases with the number of additional squares. However, a similar row consisting of two alternating shapes (squares and stars) produced a strong crowding effect. The results are naturally predicted by lateral inhibition/saliency model with a reasonable assumption that two identical items inhibit each other more heavily than different items. (Additionally, I suppose that inhibitory effect of the Vernier on two other items is negligible because of its smaller size) (Figure 6C).

Pooling at higher levels

Higher levels of the network pool over very large receptive fields. Without focused attention, effective pooling zones must be much larger than 0.5E (up to the full visual field). Global pooling can be studied by visual search (Shaw, 1980; Palmer, Ames, & Lindsey, 1993; Palmer, Verghese, & Pavel 2000), or whole report experiments (e.g., Kyllingsbaek, Valla, Vanrie, & Bundesen, 2007). In usual search task, all stimuli in a display are relevant. The results of search experiments should reveal, which kind of information can be pooled, and which rules of pooling are used at highest levels of vision.

In a recent article (Põder, 2017), I analyzed combined effects of crowding and set size in visual search experiments (Figure 7). A simple multiplicative model well describes the results, consistent with independent effects of different levels of processing. However, there is some correlation between two effects, across studied features and conjunctions of features. This could be

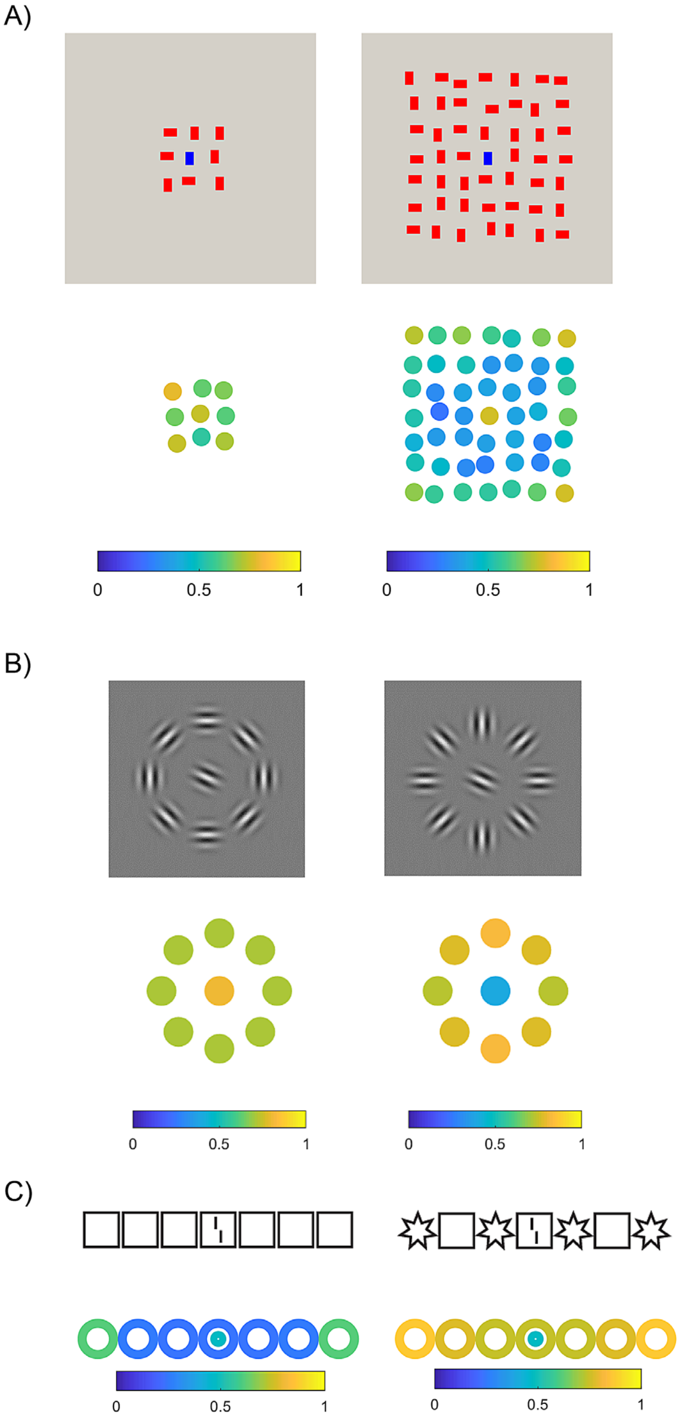

Figure 6. Examples of saliency maps based on simple lateral inhibition rules that predict the effects of (A) number of flankers (Põder, 2006), (B) configuration of orientations (Livne & Sagi, 2007), and (C) shape homogeneity/heterogeneity (Manassi et al., 2016).

expected because features in higher levels are based on the features from preceding levels. However, more detailed studies of features and pooling mechanisms over multiple levels are required. I suggest that regarding global/central capacity limitations, local lateral inhibition, and crowding effects as different instances of spatial pooling within a multilevel neural network may help to move toward a general theory of computations in vision.

Amendments to the feature integration theory

Integration of visual features by means of spatial attention (Treisman & Gelade, 1980) has been an influential theory in vision research. However, it appears that artificial neural networks can recognize objects without anything like this. Some researchers have suggested that the feature integration function of attention is unnecessary (e.g., Rosenholtz, Huang, & Ehinger, 2012). I argue that spatial attention has an important role, and, with slight modifications, the feature integration idea fits well the neural network model too.

The original feature integration model (Treisman & Gelade, 1980) supposed that “spotlight” of attention binds visual features by selecting one object at a time. Assuming minimum radius of spotlight /attentional window is 0.5*E*, it is very likely that several objects fall into it simultaneously. How will feature integration work in that case? Crowding studies suggest that human vision uses probabilistic selection of features within an attentional window, with probabilities determined by a weighting function of the window (Põder & Wagemans, 2007; Vul & Rich, 2010; Figure 8).

The classic feature integration model deals with conjunctions of different features having the same spatial positions (Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989). However, the perception of feature configurations, or relative position of features, is likely the most attention-dependent task for human vision (Bergen & Julesz, 1983; Wolfe, 1998a; Põder, 1999; Gilden, Thornton, & Marusich, 2010; Palmer, Fencsik, Flusberg, Horowitz, & Wolfe, 2011). Simple spotlight model cannot explain the role of attention in this type of tasks. However, there is a plausible way to extend the classic model (Figure 9). A small set of neurons with spatially shifted integration fields can encode position of a given feature within an area covered by these receptive fields (e.g., Baldi & Heiligenberg, 1988). These coarse feature-position signals carry some position information even when pooled over the full visual field. Relative position of two features can be decoded when there are no irrelevant similar features in a display, or when signals from irrelevant features are excluded by spatial attention.

According to the traditional view, there is a set of “simple” visual features that are registered

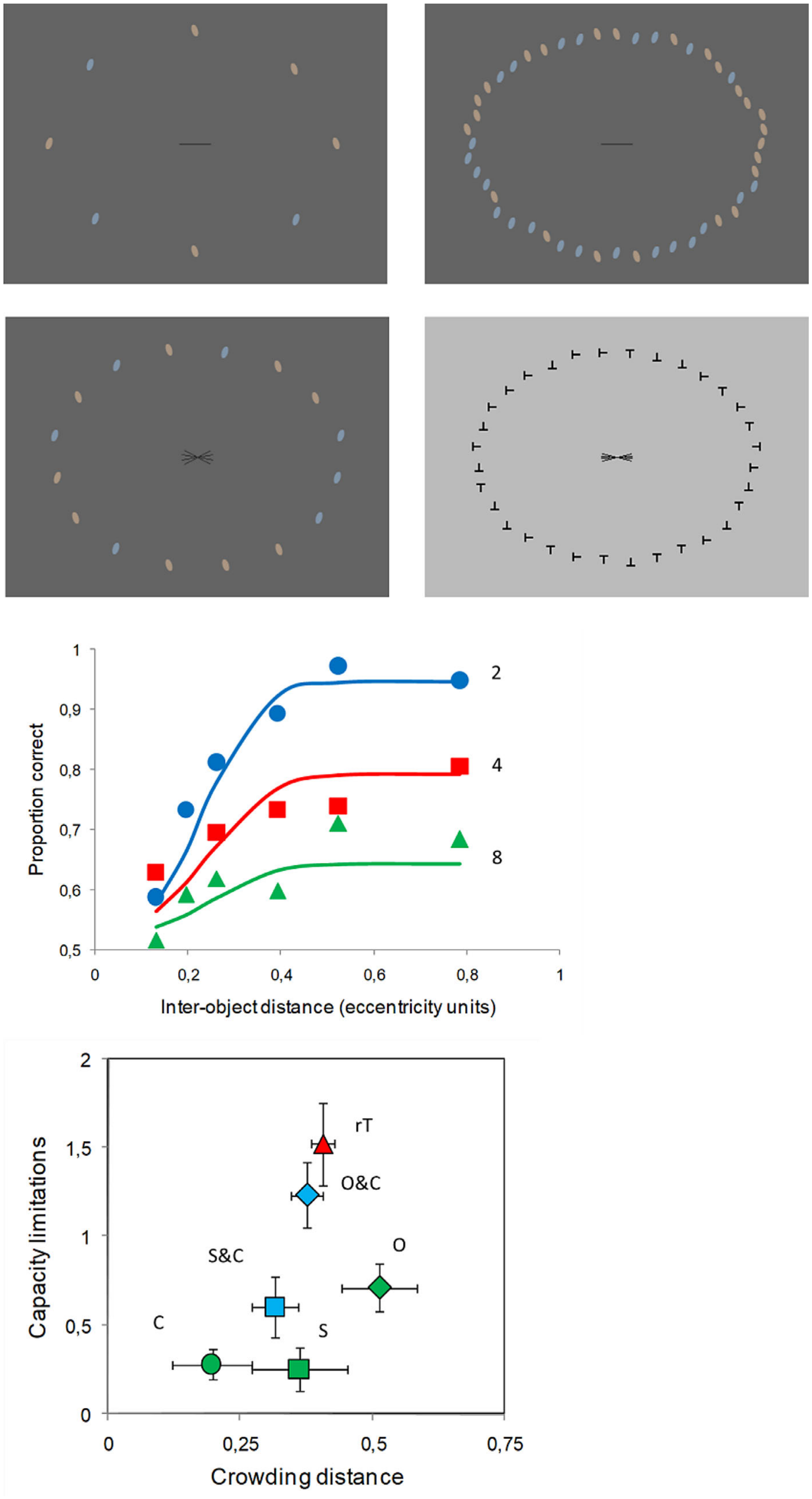

Figure 7. Combined efffects of local and global interference (Põder, 2017). Examples of stimuli where interitem distance and set size were varied (top). Example of the results of an individual experiment (middle). Symbols depict data and lines are model fits. Set sizes: 2 = blue circles, 4 = red rectangles, and 8 = green triangles. Results plotted in the space of global capacity limitations and crowding

→

←
distance (bottom). Green markers represent simple features (S, size; O, orientation; C, color), blue markers = conjunctions, and red markers = rotated Ts.

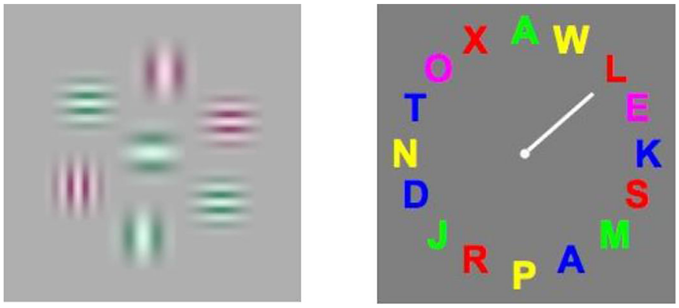

Figure 8. Reporting a target surrounded by flankers with different features reveals independent sampling of features within an attentional window. Displays from Põder and Wagemans, 2007 (left; target in the center of a group), and Vul and Rich, 2010 (right; target indicated by spatial cue; reprinted by permission of SAGE Publications Inc).

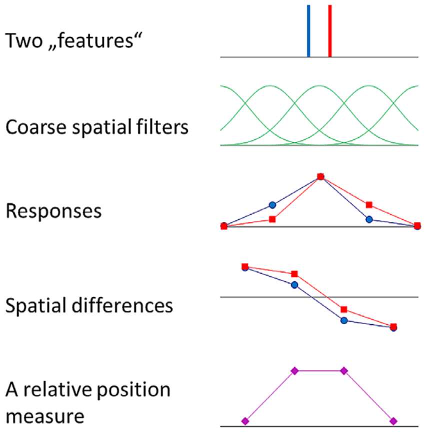

Figure 9. A hypothetical mechanism of encoding spatial positions and decoding of relative positions. Relative position information is available over a wide area, but similar features within the attentional window will make it ambiguous.

automatically/preattentively in every location of the visual field. The exact number of simple features is unknown, but usually, 10 to 20 candidates have been listed (e.g., Wolfe, 1998b).

However, we can, without focused attention, perceive quite complex differences in visual images, including statistical properties of feature distributions, or “gist” of scenes (Ariely, 2001; Torralba & Oliva, 2003). Portilla and Simoncelli (2000) had to adjust about 700 “features” (mostly spatial correlations of wavelets—a kind of simple features) to make two images indiscriminable without focused attention. Although a part of these complex features could be redundant or unnecessary for peripheral vision (Balas, 2006; Ackermann & Landy, 2014), additional features could be needed to account for some results (Wallis, Funke, Ecker, Gatys, Wichmann, & Bethge, 2019). Anyway, the number and complexity of preattentive features are likely much larger than supposed in early studies.

Conclusions

In this study, I attempted to combine the main psychophysical findings on visual crowding, attention, and central capacity limitations with hierarchical neural network model. I suggest that the similar principles of pooling and selection, at various levels of visual processing, can explain different psychophysical phenomena—visual crowding and central capacity limitations. An important factor seems to be the level where spatial attention is applied. Regardless of similar principles, the exact computations may vary across levels. For example, different pooling rules (averaging, max, correlation) may dominate at different levels of processing.

Recent crowding studies have reported many results that contradict simple pooling over fixed receptive fields. Different explanations from qualitative grouping account to relatively complex computational models have been proposed. I believe that a notion of saliency could be useful here. There is a lot of evidence for similarity-dependent lateral inhibition at different levels of biological visual systems. Similar calculations have been used in machine vision as well (e.g., Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009). Several simple cases from crowding experiments apparently fit well to this simple model. In more complex cases, saliency can be calculated at several levels of visual processing, and candidate objects at different levels may compete for access to further processing.

The present study suggests how classic ideas of attention and feature integration could be related to modern neural network models of vision. I suppose that the “simple” features to be combined should be a bit more complex, and the “spotlight” of attention has a minimum radius of about half of eccentricity. The amendments do not contradict the majority of earlier results because complex models can be reduced to simplified versions when traditional simple stimuli are used.

Keywords: visual crowding, attention, saliency, capacity limitations, feature integration, neural networks

Acknowledgments

The main points of this study were presented at the Workshop on Visual Crowding, June 23–24, 2019, Murten, Switzerland. I thank the organizers of this event.

Commercial relationships: none.

Corresponding author: Endel Põder.
Email: endel.poder@ut.ee.
Address: Institute of Psychology, University of Tartu, Näituse 2, 50409 Tartu, Estonia.

References

Ackermann, J. F., & Landy, M. S. (2014). Statistical templates for visual search. *Journal of Vision, 14*(3):18, 1–17.

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research, 16*, 71–78.

Ariely, D (2001) Seeing sets: representation by statistical properties. *Psychological Science, 12*, 157–162

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research*, *46*, 299–309.

Balas, B. J., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12):13, 1–18.

Baldi, P., & Heiligenberg, W. (1988). How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers. *Biological Cybernetics, 59*, 313–318.

Bergen, J. R., & Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. *Nature, 303,* 696–698.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*, 177–178.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology, 88*, 2547–2556.

Chen, F., Roig, G., Isik, X., Boix, L., & Poggio, T. (2017) Eccentricity dependent deep neural networks: Modeling invariance in human vision. Retrieved from http://hdl.handle.net/1721.1/112279.

Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., & Herzog, M. H. (2019). Beyond Bouma's window: How to explain global aspects of crowding? *PLoS Computational Biology, 15*(5), e1006580.

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review, 124*(4), 483.

Freeman, J., & Simoncelli, E.P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*, 1195–1201.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202.

Gilden, D.L., Thornton, T.L., & Marusich, L.R. (2010). The serial process in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 533–542.

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences, USA, 106*(31), 13130–13135.

Gross, C. G., Rocha-Miranda, C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology, 35*, 96–111.

Han, Y., Roig, G., Geiger, G., & Poggio, T. (2020). Scale and translation-invariance for novel objects in human vision. *Scientific reports*, *10*(1), 1411.

Harrison, W. J., & Bex, P. J. (2017). Visual crowding is a combination of an increase of positional uncertainty, source confusion, and featural averaging. *Scientific Reports, 7*, 45551.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision, 15*(6), 5, doi:10.1167/15.6.5.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology, 28*, 229–289.

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

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489–1506.

Jarrett, K., Kavukcuoglu, K., Ranzato, MA., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, 2146–2153.

Kooi, F., Toet, A., Tripathy, S., & Levi, D. (1994). The effect of similarity and duration on spatial interaction in peripheral-vision. *Spatial Vision, 8*(2), 255–279.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016) Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol., 12*(4), e1004896.

Kyllingsbaek, S., Valla, C., Vanrie, J., & Bundesen, C. (2007). Effects of spatial separation between stimuli in whole report from brief visual displays. *Perception & Psychophysics, 69*(6), 1040–1050, doi:10.3758/BF03193942.

Knierim, J. J., & van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology, 67*, 961–980.

Nothdurft, H. C., Gallant, J. L., & Van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: Correlates of "popout" under anesthesia. *Visual Neuroscience, 16*, 15–34.

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654.

Li, W., Their, P., & Wehrhahn, C. (2000). Contextual influence on orientation discrimination of humans and responses of neurons in V1 of alert monkeys. *Journal of Neurophysiology, 83*, 941–954.

Li, Z. (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proceedings of National Academy of Science, USA, 96*, 10530–10535.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*, 9–16.

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision, 7*(2):4, 1–12.

Livne, T., & Sagi, D. (2010). How do flankers' relations affect crowding? *Journal of Vision, 10*(3):1, 1–14,

Lonnqvist, B., Clarke, A. D. F, & Chakravarthi, R. (2020). Crowding in humans is unlike that in convolutional neural networks. *Neural Networks, 126*, 262–274.

Malania, M., Herzog, M. H., & Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *Journal of Vision, 7*(2), art: 1.

Manassi, M., Lonchampt, S., Clarke, A., & Herzog, M. H. (2016). What crowding can tell us about object representations. *Journal of Vision, 16*(3):35, 1–13, doi:10.1167/16.3.35.

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision, 13*(13), art: 10

Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology, 28*(3), R127–R133.

Motter, B. C. (2018). Stimulus conflation and tuning selectivity in V4 neurons: A model of visual crowding. *Journal of Vision, 18*(1), 15, doi:10.1167/18.1.15.

Nothdurft, H. C., Gallant, J. L., & Van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: Correlates of "popout" under anesthesia. *Visual Neuroscience, 16*, 15–34.

Palmer, E. M., Fencsik, D.E., Flusberg, S.J., Horowitz, T.S., & Wolfe, J.M. (2011). Signal detection evidence for limited capacity in visual search. *Attention, Perception and Psychophysics, 73*(8):2413–24.

Palmer, J., Ames, C. T., & Lindsey, D. T. (1993). Measuring the effect of attention on simple visual search. *Journal of Experimental Psychology: Human, Perception and Performance, 19*, 108–130.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*, 1227–1268.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature detection and integration. *Journal of Vision, 4*, 1136–1169.

Pelli, D., & Tillman, K. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*(10), 1129–1135.

Portilla, J., & Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*, 49–71.

Põder, E. (1999). Search for feature and for relative position: Measurement of capacity limitations. *Vision Research, 39*, 1321–1327.

Põder, E. (2006). Crowding, feature integration, and two kinds of "attention". *Journal of Vision, 6*(2), 163–169.

Põder, E. (2007). Effect of colour pop-out on the recognition of letters in crowding conditions. *Psychological Research, 71*(6), 641–645.

Põder, E. (2017). Combining local and global limitations of visual search. *Journal of Vision, 17*(4):10, 1–12.

Põder, E., & Wagemans, J. (2007). Crowding with conjunctions of simple features. *Journal of Vision, 7*(2):23, 1–12.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *3*, 1199–1204.

Rolls, E. T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron, 27*, 1–20.

Rosen, S., & Pelli, D. G. (2015). Crowding by a repeating pattern. *Journal of Vision, 15*(6), 10, doi:10.1167/15.6.10.

Rosenholtz, R., Huang, J., & Ehinger, K.A. (2012). Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3*, 13, 1–15.

Rosenholtz, R., Yu, D., & Keshvari, S. (2019) Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of Vision. 19*(7):15, 1–25.

Saarela, T.P., Sayim, B., Westheimer, G., & Herzog, M.H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, *9*(2):5, 1–11.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature, 378*, 492–496.

Shaw, M. L. (1980). Identifying attentional and decision-making components in information processing. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 277–296). Hillsdale, NJ: Erlbaum.

Strasburger, H., & Malania, M. (2013). Source confusion is a major cause of crowding. *Journal of Vision, 13*(1), 24–24.

Strasburger, H. (2007). Unfocussed spatial attention underlies the crowding effect in indirect form vision. (vol 5, pg 1024, 2005). *Journal of Vision, 7*(3), 7, doi:10.1167/7.3.7.

Strasburger, H., Rentschler, I., & Juttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision, 11*(5), 13.

Torralba, A., & Oliva, A. (2003). Statistics of natural images categories. *Network: Computation in Neural Systems, 14*, 391–412.

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

Van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLoS Computational Biology, 6*(1), e1000646.

Vickery, T. J., Shim, W. M., Chakravarthi, R., Jiang, Y. V., & Luedeman, R. (2009). Supercrowding: Weakly masking a target expands the range of crowding. *Journal of Vision, 9*(2), 12.1–15, doi:10.1167/9.2.12.

Vul, E., & Rich, A. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science, 21*(8), 1168–1175.

Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2019). Image content is more important than Bouma's Law for scene metamers. *eLife, 8*, e42512, 1–43.

Wolfe, J. M. (1998a) What can 1 million trials tell us about visual search? *Psychological Science. 9*(1), 33–39.

Wolfe, J. M. (1998b). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13–73). Philadelphia: Psychology Press.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration theory of attention. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 419–433.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7):32, 1–20.

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T. A., & Desimone, R. (2011) Object decoding with attention in inferior temporal cortex. *Proceedings of National Academy of Science, USA, 108*(21), 8850–8855.