RESEARCH ARTICLE

Shrinking Bouma's window: How to model crowding in dense displays

Alban Bornet[1☯*], **Adrien Doerig**[1,2☯], **Michael H. Herzog**[1], **Gregory Francis**[3], **Erik Van der Burg**[4,5]

[1] Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [2] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands, [3] Department of Psychological Sciences, Purdue University, West Lafayette, Indiana, United States of America, [4] TNO, Human Factors, Soesterberg, The Netherlands, [5] Brain and Cognition, University of Amsterdam, Amsterdam, The Netherlands

☯ These authors contributed equally to this work.
* alban.bornet@epfl.ch

Abstract

In crowding, perception of a target deteriorates in the presence of nearby flankers. Traditionally, it is thought that visual crowding obeys Bouma's law, i.e., all elements within a certain distance interfere with the target, and that adding more elements always leads to stronger crowding. Crowding is predominantly studied using sparse displays (a target surrounded by a few flankers). However, many studies have shown that this approach leads to wrong conclusions about human vision. Van der Burg and colleagues proposed a paradigm to measure crowding in dense displays using genetic algorithms. Displays were selected and combined over several generations to maximize human performance. In contrast to Bouma's law, only the target's nearest neighbours affected performance. Here, we tested various models to explain these results. We used the same genetic algorithm, but instead of selecting displays based on human performance we selected displays based on the model's outputs. We found that all models based on the traditional feedforward pooling framework of vision were unable to reproduce human behaviour. In contrast, all models involving a dedicated grouping stage explained the results successfully. We show how traditional models can be improved by adding a grouping stage.

Author summary

To understand human vision, psychophysical research usually focuses on simple stimuli. Vision is often described as a cascade of feed-forward computations in which local feature detectors pool information along the processing hierarchy to form complex and abstract features. Crowding can be modelled within this framework by the pooling of information from one processing stage to the next. This naturally explains Bouma's law, a hallmark of crowding according to which only elements within a certain region, often proposed to be half the target eccentricity, interfere with the target. However, pooling models are strongly challenged by recent experimental results, because Bouma's law does not hold for more complex stimuli. Visual elements far beyond Bouma's window can increase or alleviate

OPEN ACCESS

Citation: Bornet A, Doerig A, Herzog MH, Francis G, Van der Burg E (2021) Shrinking Bouma's window: How to model crowding in dense displays. PLoS Comput Biol 17(7): e1009187. https://doi.org/10.1371/journal.pcbi.1009187

Editor: Wolfgang Einhäuser, Technische Universitat Chemnitz, GERMANY

Received: October 22, 2020

Accepted: June 16, 2021

Published: July 6, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1009187

Data Availability Statement: The code to reproduce all models' results and the data of the human experiment are available at https://bitbucket.org/albornet/shrinking_boumas_window.

Funding: AB was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 785907 (Human Brain Project SGA2) and No. 945539 (Human Brain Project SGA3). AD was supported by the Swiss National Science Foundation grants n.176153 "Basics of visual processing: from elements to figures" and n.191718 "Towards machines that see like us: human eye movements for robust deep recurrent neural networks". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

crowding. In addition, Van der Burg and colleagues showed that only the nearest neighbours interfere with the target in dense displays. Hence, Bouma's window can shrink too. Here, we aimed at modelling the range of crowding in dense displays. From previous studies, we know that visual crowding cannot be explained without grouping and segmentation. We compared the performance of different models of vision to the human data of Van der Burg and colleagues. We found that all models based on the traditional pooling framework of vision failed to reproduce the human data, whereas all models that included grouping and segmentation processes were successful in this respect. We concluded that grouping and segmentation processes naturally and consistently explain the difference between simple and complex displays in vision paradigms.

Introduction

In the classic framework, vision is a feed-forward process that starts with the analysis of basic features such as oriented edges [1–4]. These basic features are pooled along the visual hierarchy to form more complex feature detectors, until neurons respond to objects [5–9]. A strength of modelling visual perception as such a feedforward process is that it breaks down the complexity of vision into mathematically tractable sub-problems. However, it has become clear that this classic framework cannot account for a wide range of experimental results [10–16].

For example, in a vernier discrimination task, two slightly offset vertical bars are presented in the periphery of the visual field (Fig 1A). The task is to determine whether the bottom bar is offset to the left or to the right. The task is easy when the target is displayed in isolation (Fig 1B, red dashed line). Adding a square around the vernier severely impairs performance (i.e., visual crowding, Fig 1B, first column).

In the classic framework, such impairments are explained by flankers and target features being pooled along the visual hierarchy [17–20]. For example, in Fig 1C, the vernier target and the flankers are pooled, which deteriorates the representation of the vernier. It is often claimed that: a) only elements within the pooling distance, i.e., inside the so-called Bouma's window (equal to half the target eccentricity), affect each other [21–24] and b) adding more flankers within this window always leads to more crowding because more irrelevant information is pooled.

However, recent research has shown many effects that cannot be explained in this classic framework. For example, flankers far from the target (and even far outside Bouma's window) can in fact strongly *improve* performance, depending on the global configuration of the stimulus (uncrowding [10, 11, 25–31]; Fig 1B, second to last columns). As another example, it has been shown that detailed information within Bouma's window can survive crowding [32, 33]. Hence, a) interactions are *not* restricted to Bouma's window and b) adding flankers does *not* always deteriorate information.

Obviously, studies with sparse displays cannot reveal these important effects. Displays that contain a large number of flankers come with the problem that the number of configurations increases exponentially with the number of flankers. For example, a relatively simple array of 8 by 8 either vertical or horizontal flankers has more possible configurations than there are seconds since the Big Bang. Hence, it is hard to determine which configurations show interesting effects that are not captured by the classic framework of vision. How can these configurations be discovered?

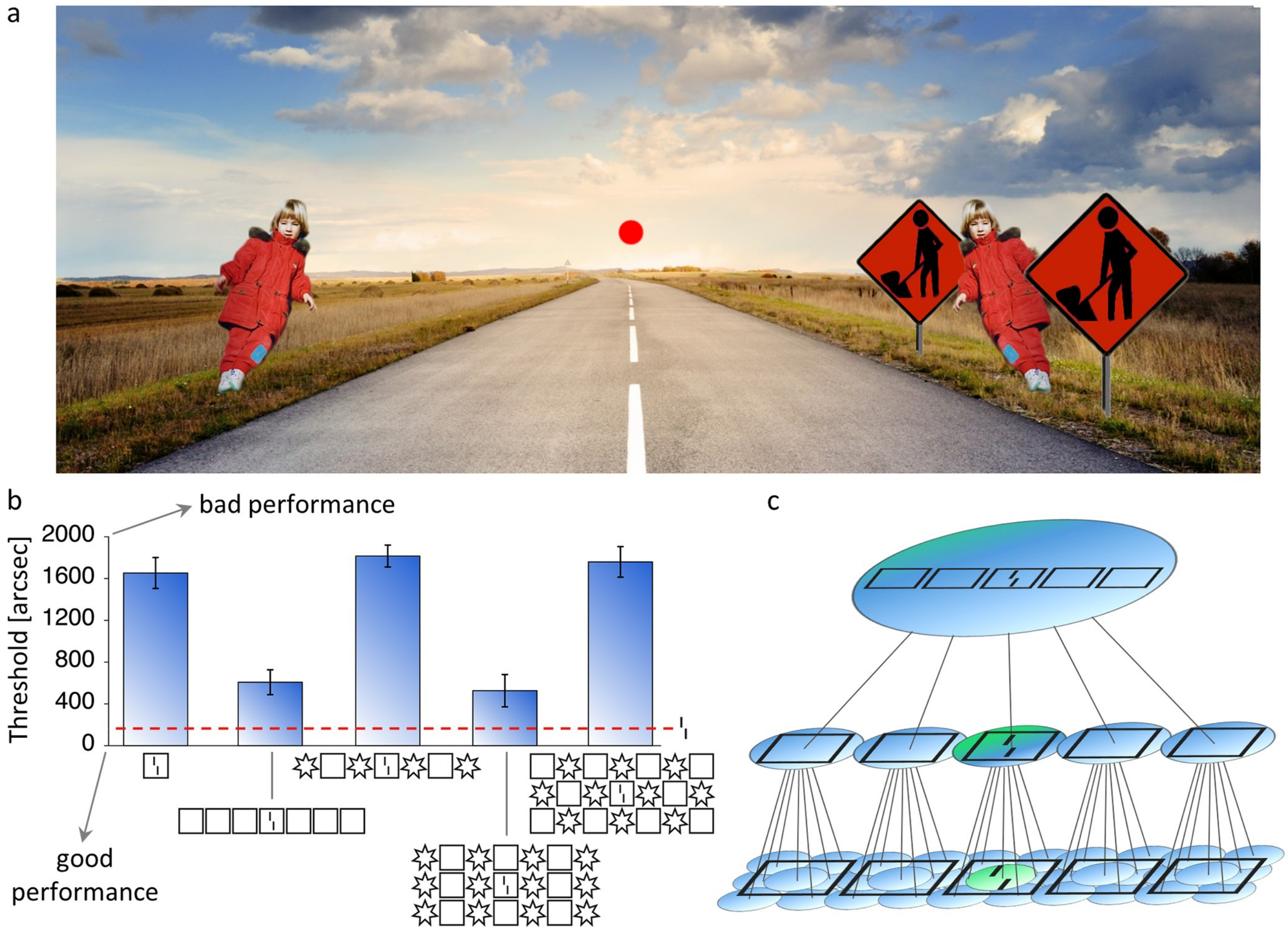

Fig 1. **a. Visual crowding in everyday life.** When looking at the red fixation dot, the child on the right is more difficult to identify than the same child on the left, because the nearby signposts lead to crowding (adapted from [34]). **b.** Manassi et al. [28] presented a vernier in the periphery, surrounded by different flanker configurations. The y-axis shows the vernier offset threshold for 75% of correct responses (the larger the threshold, the worse the performance). In the absence of flankers, the threshold is low (red dashed line). When a square is placed around the target, the task is much harder (crowding, 1st column). When more squares are added, performance recovers almost to the unflanked level (uncrowding, 2nd column). Crowding strength is strongly affected by the whole flanker configuration (3rd to last columns). **c. Classic hierarchical model of crowding.** Local information is pooled along the feedforward hierarchy of the visual system, to form more complex feature detectors. In this example, neurons (circles represent the extent of their receptive fields) detect simple oriented features in the first layer, simple shapes in the second layer and shape configurations in the last layer. Along the hierarchy, pooled activity dilutes information related to vernier offset. In this view, adding more flankers can only lead to stronger crowding. Adapted with permission from [16].

https://doi.org/10.1371/journal.pcbi.1009187.g001

Recently, Van der Burg et al. [35] proposed a paradigm in which observers had to discriminate an almost vertical target, slightly tilted to the left or to the right, embedded in different configurations of vertical and horizontal flankers. First, Bouma's law was verified using *sparse displays*, in which only 4 either vertical or horizontal flankers surrounded the target (Fig 2A). Then, they presented rectangular arrays of 15x19 bars (284 horizontal or vertical flankers and 1 tilted target; *dense displays;* Fig 2B, top). Understanding which distractors at what location interfere with target identification in dense displays is difficult (if not impossible) using a factorial design, as there are 2^{284} possible display configurations.

To circumvent this problem, Van der Burg et al. [35] used a genetic algorithm (*GA* [36]; Fig 2B, bottom). In this study, participants performed an orientation discrimination task. For each participant, the displays that led to the highest accuracy were selected and combined using a

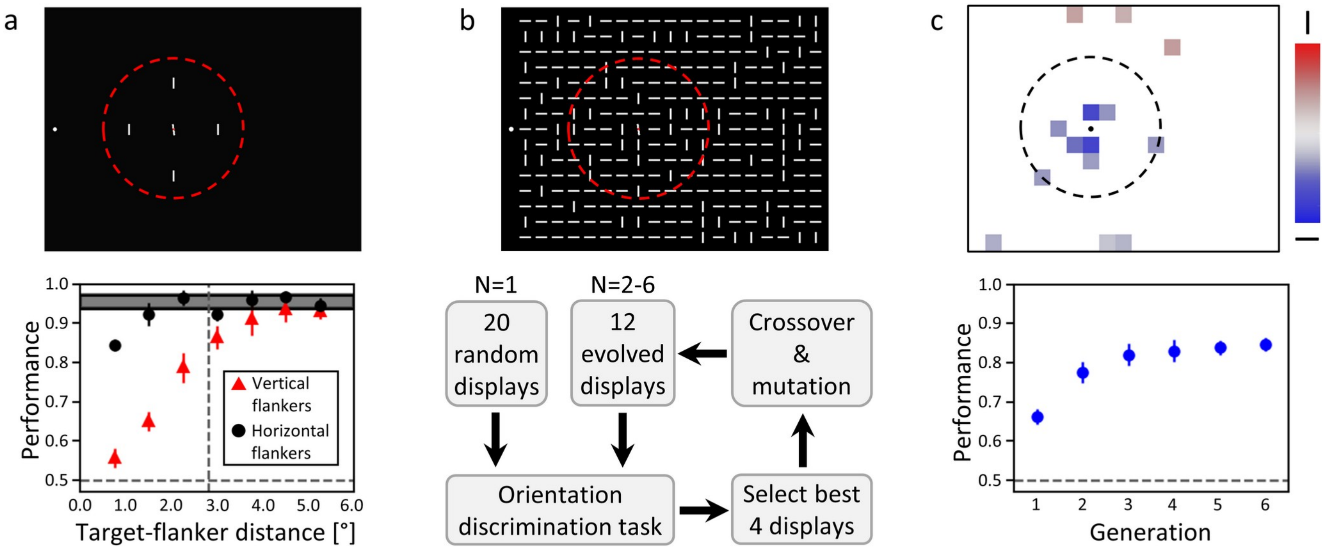

Fig 2. a. Top. Example display of the crowding experiment involving sparse displays in Van der Burg et al. [35]. Observers reported whether the target was tilted to the left or right from vertical while fixating the white dot on the left. The target was surrounded by either four horizontal or vertical flankers. The dashed circle, which was not visible during the experiment, indicates Bouma's window. **Bottom.** Human performance (proportion of correct responses) for both flanker orientations and different target-flanker distances. Error bars indicate the standard deviations across observers. The shaded area corresponds to the unflanked condition. The horizontal dashed lines indicate chance level performance. The vertical dashed line indicates Bouma's window. Less crowding was observed for horizontal flankers and Bouma's law was verified. **b. Top.** Example of a dense display. The task was the same as in the sparse display experiment. **Bottom.** GA procedure used in Van der Burg et al. [35]. For every participant, 20 dense displays (whose proportion of vertical flankers was set to lead to 67% of performance) were chosen as the first generation (N = 1). Then the displays that led to the highest accuracy were selected as the parents of the next generation (children). This selection process was repeated for 6 generations of displays (N = 2–6). **c.** Results of the GA procedure in Van der Burg et al. [35]. **Bottom.** During the GA procedure, human performance increased over generations. **Top.** Map depicting which locations in dense displays were crucial for the performance improvements caused by the GA procedure. For each flanker location, the proportion of vertical or horizontal flankers in generation 6, over all participants, was compared (two-tailed t-test) to displays coming from a random selection process between generations (neutral condition). Red/blue slots correspond to locations in which the proportion of horizontal/vertical flankers increased significantly after the evolution process ($p < 0.05$, not corrected for multiple comparisons to increase the possibility to find evidence for Bouma's law). Colour intensity represents effect size, i.e., in what proportion a vertical or horizontal flanker is selected. White spaces indicate that neither vertical nor horizontal flankers interfered with the target.

https://doi.org/10.1371/journal.pcbi.1009187.g002

crossover and mutation procedure to generate the next generation of displays. This process was repeated over six generations to maximize human performance (see Methods for more details; see [37–39] for a similar methodology to study visual search in complex displays). Using this procedure, performance increased dramatically over generations (Fig 2C, bottom). Interestingly, this improvement was predominantly caused by the target's nearest neighbours and, to a lesser extent, by other flankers within a radius of 1° (Fig 2C, top). It seems as if Bouma's window has shrunk.

Here, we investigated which models of crowding can explain these results. To do so, we applied the same GA procedure as in Van der Burg et al. [35], but instead of selecting the displays based on human performance, we selected them based on model performance. First, we tested several leading models of crowding that are based on the classic feedforward pooling framework of vision: a model that artificially reproduced Bouma's law in dense displays (*Bouma model*; see S1 Appendix), a population coding model (*Popcode* model [40]; see S2 Appendix), a model based on summary statistics (*Texture model* [41]; see S3 Appendix) and a feedforward convolutional neural network classifier (*CNN classifier* [16, 42]; see S4 Appendix).

However, we did not expect the former models to reproduce human behaviour for dense displays. Indeed, several studies found that a visual grouping stage is necessary to explain global configuration effects in crowding [15, 16, 43]. For this reason, we also tested several models of crowding that include grouping and segmentation processes: a model of low-level segmentation (*Laminart model* [44]; see S5 Appendix), a classic convolutional neural network augmented with recurrent grouping processes (*Capsule network* [43, 45]; see S6 Appendix) and a model that combined the population coding and the segmentation models (*Popart model*; see S7 Appendix).

We show that only the models that contain a dedicated grouping mechanism explain the results of Van der Burg et al. [35]. Hence, we propose that grouping is required to explain which elements *within* Bouma's window affect target discrimination performance. Because grouping is also crucial to understand which elements *beyond* Bouma's window impact performance [15], we propose that visual grouping (and not Bouma's law) determines the range of interactions in crowding and naturally and consistently explains why this range highly depends on the nature and the configuration of the visual stimulus.

Methods

Ethics statement

Participants gave oral consent before the experiment, which was conducted in accordance with the Declaration of Helsinki except for the preregistration (World Medical Organization, 2013) and was approved by the local ethics committee (Commission d'éthique du Canton de Vaud, protocol number: 164/14, title: Aspects fondamentaux de la reconnaissance des objets protocole général).

The stimuli and the GA procedures were the same as in Van der Burg et al. [35]. We simply replaced human observers with models. The displays were composed of a target (a bar tilted by either +5 or -5 degrees from vertical) embedded in a dense array of 284 flanking bars, each of which was either vertical or horizontal, positioned in a regular and rectangular grid of 15 rows and 19 columns, spanning 11.25° by 14.25° (see Fig 2B, top, for an example display). Details about how spatial units are represented in each model are given in the appendices. The fixation point (when the tested model used one) was located 0.75° to the left of the centre of the leftmost column. The target was always displayed at the same position (8th row, 8th column, eccentricity = 6°) and the task of the models was to report the target orientation (tilted to the left or to the right from vertical). As in the human experiments of Van der Burg et al. [35], model performance for each display was always computed as the proportion of correct responses in 12 trials.

For each model, the GA procedure started with 20 dense displays featuring random configurations of flankers (first generation). The 4 configurations that led to the best model performance were selected as parent configurations. Then, for each model, 12 children configurations were generated by randomly mixing the parent nodes. Each child node had a 50% chance to come from the first parent display and another 50% chance to come from the second one. After this crossover procedure, each node had a 4% chance to be randomly assigned to either a horizontal or a vertical flanker (i.e., a mutation procedure). Those new configurations constituted the next generation of the GA. The same generative process was repeated for 6 generations. To reduce noise, the whole GA was run 4 times, like in Van der Burg et al. [35], where each participant performed 4 sessions.

For each model, we monitored the proportion of vertical and horizontal flankers at each location of the dense displays in the last generation and compared all of them to the respective proportions in the last generation of a random selection process, i.e., a neutral condition, as in Van der Burg et al. [35]. In this neutral condition, the GA parameters were the same as when running the models, except that the displays were selected randomly between the generations. The difference between the model behaviour and the random selection behaviour is presented as a map where a red or a blue slot respectively indicate that the GA procedure selected a significantly larger fraction of vertical or horizontal flankers at that location, compared to the last generation of randomly selected displays (two-tailed t-tests; $p < 0.05$). Like in Van der Burg et al. [35], the statistical tests were not corrected for multiple comparisons to maximize the possibility of finding evidence for Bouma's law in the results. We call this the *selection measure* (see Fig 2C, top, for corresponding human results). In addition, we made sure that the GA procedure worked, i.e., that model performance increased over generations. We call this the *performance measure* (see Fig 2C, bottom, for corresponding human results). In the results section, we refer to both the performance and the selection measures as the *GA measures*.

In the GA procedure of Van der Burg et al. [35], the proportion of vertical flankers in the first generation of dense displays was set to lead to an initial performance of 67% for each human observer to avoid floor and ceiling effects. Here, we wanted to make a fair comparison between different models. If two models would require for example 10% and 90% of vertical bars, respectively, to have a performance of 67% in the first generation of displays, it would be easier to see a significant increase of horizontal bars in the subsequent generations for the second model than for the first one. For this reason, the initial proportion of vertical flankers was set to a single value for all models, which corresponds to the mean of what was used in Van der Burg et al. [35], i.e., 30% of vertical flankers in the first generation. Prior to the GA procedure, we tuned the parameters of each model to obtain a performance of 67% in dense displays with 30% of vertical flankers. The goal was to find the best parameters for an optimal GA procedure and to have the fairest comparison between models. The performance of some models was bounded by a value lower than 67%. Only in these cases, we adapted the target orientation amplitude so that higher performances than 67% could be reached, thereby allowing the model parameters to be tuned to the required level of performance.

Moreover, we tried as much as possible to tune the model parameters to reproduce Bouma's law in sparse displays. The reason is that our main question was whether the models we tested could reproduce Bouma's law in sparse displays while shrinking their pooling range in dense displays. Hence, we were not interested in models with a small pooling range that could easily reproduce human behaviour in the selection measures, but not Bouma's law in sparse displays. To this end, we measured model performance for the same sparse display experiment as in Van der Burg et al. [35], in which Bouma's law was observed. We call this the *sparse display measure* (see Fig 2A, bottom, for corresponding human results). Finally, we assessed the different models' behaviour for randomly generated dense displays in which the proportion of vertical flankers varied from 0.0 to 1.0 by increments of 0.2. We call this the *proportion measure*. Note that we performed the proportion measure with humans as well, because this experiment was not conducted by Van der Burg et al. [35] (see S8 Appendix for more details).

The GA measures are reported in the Results section by running each model 10 times, to simulate 10 different human subjects, as in the GA procedure of Van der Burg et al. [35]. The reported standard deviations are computed over these 10 runs. The sparse display and proportion measures are not based on statistical testing (i.e., they are not part of the GA procedure) and are hence reported by running each model 100 times, to obtain clearer results. The code for the entire procedure is available at https://bitbucket.org/albornet/shrinking_boumas_window, as well as the code for the different models we tested and instructions to test any other model with the GA procedure.

Results

Results for all models are summarized in Fig 3. Specific descriptions of the models and details about the results can be found in the supporting information.

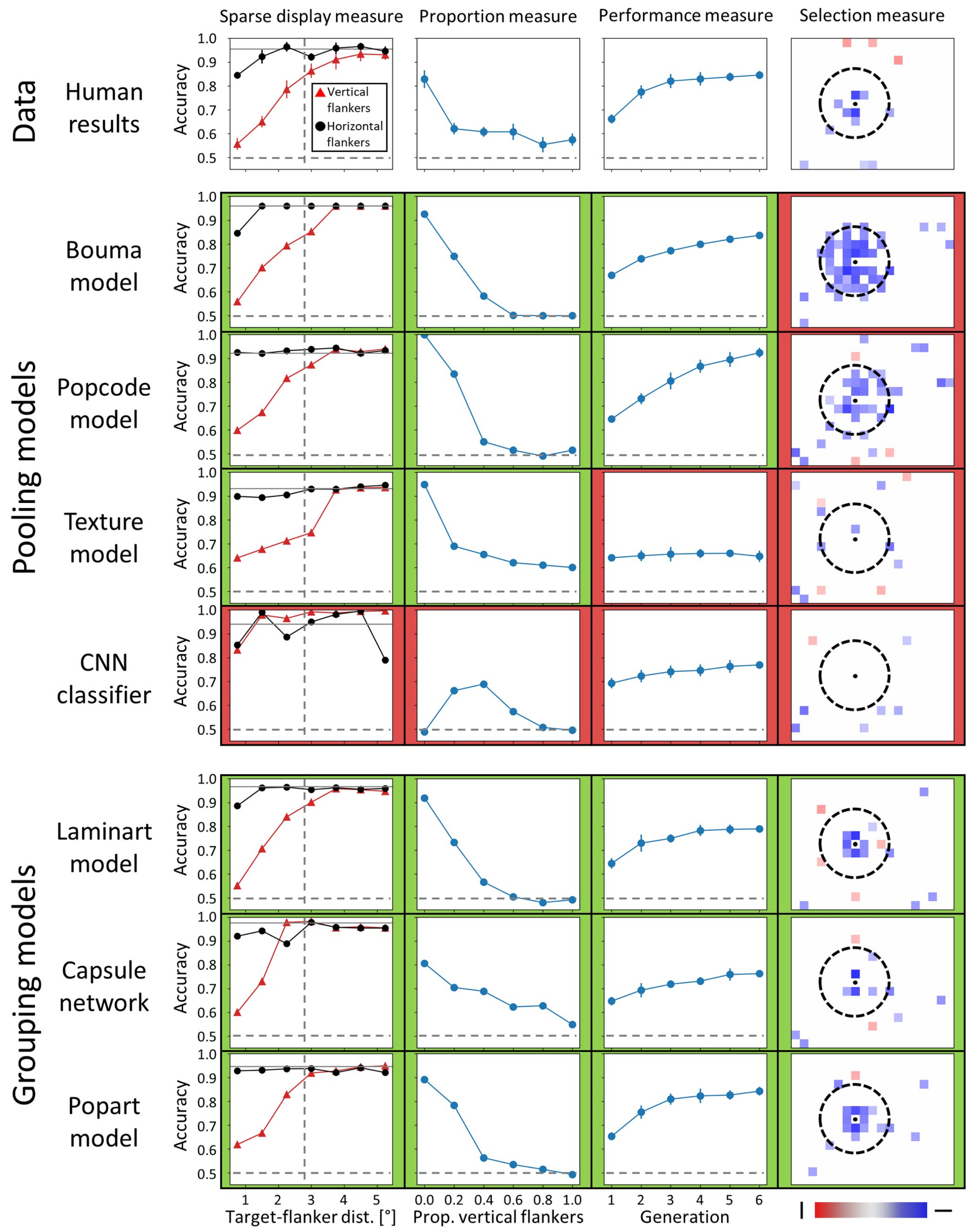
Sparse display measure
Proportion measure
Performance measure
Selection measure
Data
Human results
Pooling models
Bouma model
Popcode model
Texture model
CNN classifier
Grouping models
Laminart model
Capsule network
Popart model
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
Vertical flankers
Horizontal flankers
1
2
3
4
5
Target-flanker dist. [°]
0.0
0.2
0.4
0.6
0.8
1.0
Prop. vertical flankers
1
2
3
4
5
6
Generation

Fig 3. Results for all models, for the four measures described in the Methods section. The first row contains the human data. Every column contains a different measure of the models' behaviour and can be compared to the corresponding human data. Each measure is described in detail in the Methods section. For every measure and every model, green/red frames indicate whether a model did/did not qualitatively reproduce the corresponding human data, respectively. For the performance measure, green corresponds to an improvement of at least 10 points of accuracy during the GA procedure. For the other measures, green corresponds to a similar shape in the distribution of the model results and the human data. Note that a quantitative measurement of the similarity between the model results and the human data can be found in S10 Appendix. The vertical dashed lines in the sparse display measure and the dashed circles in the selection measure indicate the limit of Bouma's window. The horizontal dashed lines in all measures indicate chance level accuracy. In general, all models were able to reproduce the sparse display measure and the proportion measure, except for the CNN classifier. Moreover, all models based on the traditional, feed-forward pooling framework of vision failed to reproduce human results for the GA measures (performance and selection measures), either because the GA procedure was unable to find flanker configurations that improved model's performance (Texture model, CNN classifier) or because too many elements within Bouma's window were highlighted by the GA procedure (Bouma model, Popcode model). Finally, all models that contain a grouping stage qualitatively reproduced human results for the GA measures. Note that we included a fine-grained version of the selection measures in S11 Appendix, i.e., in which the fraction of vertical or horizontal flankers selected by the GA procedure is shown for all locations, regardless of whether the differences to the random GA selection process are statistically significant or not.

https://doi.org/10.1371/journal.pcbi.1009187.g003

Pooling models

First, to rule out the possibility that the GA procedure itself produced the shrinking of Bouma's window, we repeated what was done in Van der Burg et al. [35] and used a simple linear pooling model whose weights were fitted to produce Bouma's law (*Bouma model*). The model qualitatively reproduced the human data for the proportion measure and the sparse display measure but failed to reproduce the human GA measures Fig 3, second row), suggesting that the GA procedure does not produce the shrinking of Bouma's window by itself.

Then, we tested more advanced models based on the traditional, feed-forward pooling framework of vision. First, we used a model based on the population coding idea (*Popcode model* [40]). This model provides a physiologically plausible description of feature integration that accounts for various fundamental features of crowding. Second, we used a model of texture computation (*Texture model* [41]), based on low-level summary statistics, which can be seen as high-dimensional pooling [19]. Texture models may be particularly well suited for dense displays, because they encode complex natural information in a very efficient manner. Third, we used a deep convolutional neural network (*CNN classifier* [42]). Deep neural networks can be seen as a chain of nested pooling and convolution operations. They contain millions of parameters from which unexpected behaviours could arise. The results obtained with these pooling models are shown in Fig 3 (3rd to 5th rows). Except for the CNN classifier, all pooling models qualitatively reproduced human results for the sparse display and the proportion measures. However, they all failed to reproduce human data for the GA measures, either because no specific configuration was found by the GA procedure to steadily increase model performance (Texture model, CNN classifier) or because too many elements within Bouma's window were highlighted by the GA procedure (Popcode model, Bouma model). More details are in Appendices S1 to S4.

Grouping models

Finally, we tested several models that describe vision as a two-stage process. In such models, prior to interference such as depicted in the former models, visual elements are parsed into different perceptual groups. Interference only happens after the grouping stage and hence only occurs within these groups. First, we used a model of segmentation based on the recurrent integration of low-level contours (*Laminart model* [44]). The interference stage is the same as in the Bouma model. Second, we used a *Capsule Network*, a type of deep neural convolutional network that includes recurrent processing to implement grouping and segmentation [43, 45]. The results obtained with these models are shown in Fig 3 (6th and 7th rows). Both models qualitatively reproduced the human results for the sparse display and the proportion measures. Importantly, both models were also able to qualitatively reproduce the human results for the

GA measures: the radius for target-flanker interaction shrank to the nearest neighbour distance.

Despite their success at explaining the shrinking of Bouma's window, these two-stage models face problems of their own. Interference in the Laminart model was fitted to the human sparse measure data, and the Capsule network is difficult to train properly (see S5 and S6 Appendices for details). Exploiting the strengths of visual grouping and of a sophisticated interference mechanism, we combined the Laminart and the population coding models, to test if such a combination would lead to a happy marriage between both families of models (*Popart model*). Indeed, this combined model was able to reproduce human behaviour for all measures (Fig 3, last row; see S7 Appendix for more details).

We also included control simulations in which we tuned the parameters of the pooling models to reproduce human behaviour in the selection measure (instead of the sparse display measure) and checked how they behaved in sparse displays (see S9 Appendix). All pooling models include a parameter that defines their pooling range, which was modified to shrink interactions to the nearest neighbour distance. We show that tuning these models to shrink Bouma's window in dense displays prevents them from reproducing Bouma's law in sparse displays. Hence, only the grouping models can produce a small interaction range in dense displays, while keeping a Bouma-sized range in sparse displays.

Discussion

To understand crowding and vision in general, paradigms with few elements are the choice to control for complexity and unwanted interactions. For example, based on the traditional framework of vision, many studies have investigated crowding with a target and only a few flankers with a focus on local interference [17–20]. However, such simple paradigms may lead to carved-in-stone principles that are true only in such simple cases but do not apply to realistic situations. As shown here and in many previous publications, this problem seems to manifest in crowding. For example, Bouma's law holds true only for sparse displays [27, 28, 35, 46]. However, complex displays come with their own problems, which are absent in sparse displays. For example, with many flankers, the question is not only *how* visual elements interfere with the target, which is the main question in almost all crowding studies, but also *which* elements interact with each other. In addition, it is difficult to determine which displays to test out of the virtually infinitely many possible ones. To cope with the latter problem, Van der Burg et al. [35] proposed to use a GA procedure to study crowding in dense displays. In their paradigm, among all elements within Bouma's window, only the target's nearest neighbours had an influence on target discrimination performance. Importantly, the shrinking of Bouma's window in dense displays cannot be explained by the large flanker array providing a spatial cue towards the target's location. Indeed, the target is not in the centre of the flanker array. It is at the 8th row and 8th column of a 15 rows by 19 columns flanker array.

Here, we applied the GA procedure to many different models of visual crowding, each coming with its specific hypotheses about the visual system. Such an extensive comparison is a good way to test general principles of vision, because it is possible to identify, among all models, the common causes for the failure or success to explain the results. We have shown that none of the tested models that are based on a cascade of feedforward computations and pooling are able to reproduce the findings of Van der Burg et al. [35]. These models produced results in which either no element or too many elements within Bouma's window were found to interfere. In contrast, all models that include a grouping process could reproduce the human results. It seems that a global grouping and segmentation process is crucial to explain crowding in dense displays. Importantly, the combination of a global grouping stage, implemented by the Laminart model, and a local interference stage, implemented by the Popcode model, matched human behaviour in sparse as well as in dense displays (Popart model), suggesting that a happy marriage is possible between grouping and pooling models.

Of course, many other models could potentially address these results. For example, we could train a feedforward neural network to only consider the nearest neighbour flankers, since feedforward networks are universal function approximators [47]. However, such a model would be scientifically stale, because crowding is better seen as a probe into visual processing rather than as an explanatory goal per se. A model that only explains this paradigm is useless. For example, feedforward models tuned to explain the shrinking of Bouma's window in the selection measure do not reproduce Bouma's law in sparse displays (see S9 Appendix). The goal is not to overfit on a particular paradigm, but to test how processing characteristics of different hypotheses generalize to this new particular paradigm (crowding in dense displays). CNNs reach human level performance on various complex visual tasks and are subject to crowding. Summary statistics models can explain how humans process complex images without undergoing cognitive overload and capture many characteristics of visual crowding. Segmentation processes are important to solve ill-posed problems of vision and capture the effects of flankers that lie beyond Bouma's window in crowding (e.g. uncrowding). Each of these modelling frameworks has been fruitful in other areas and the goal is to test how they *generalize* to crowding in dense displays, to uncover strengths and weaknesses of each approach.

Along the same lines, Manassi et al. [28] showed that elements beyond Bouma's window can have a strong impact on target discrimination, and that the configuration of elements in the whole visual field determines crowding strength (see also [26, 27]). A similar extensive comparison of models showed, once again, that only models that could reproduce these results contained a dedicated grouping stage [15] (see also [16, 43, 48]). Moreover, Van der Burg et al. [49] showed that crowding in dense displays does not depend on target eccentricity but only on the configuration of the nearest neighbours. The grouping models that we tested here can exhibit uncrowding at the same time as the shrinking of Bouma's window, depending on the specific configuration of the flankers. In contrast, a model that modulates the window of integration based on the number of flankers but not their configuration, such as divisive normalization [50], will not be able to explain why faraway elements interact only in certain cases (e.g. why Manassi and colleagues found very long ranging interactions but Van der Burg and colleagues found the opposite). For all these reasons, it becomes clear that grouping, and not Bouma's window, determines which elements interfere with each other in human vision. In summary, our results do not prove that grouping and segmentation processes are the only way to shrink Bouma's window in dense displays, but rather show that they are the best at explaining crowding overall.

There are many more architectures for feedforward CNNs, such as ResNet and VGG. We think that these networks face similar problems as the feedforward CNNs tested here because they are also based on pooling. Because of this pooling, performance always deteriorates when flankers are added, irrespective of the global configuration of elements (for an in-depth argument, see Doerig et al. [16]). In support of this claim, neither AlexNet (see S4 Appendix) nor the capsule networks controls (see S6 Appendix, feedforward and recurrent CNNs) can explain the shrinking of Bouma's window. In addition, Geirhos et al. [51] showed that CNNs are remarkably consistent with one another behaviourally, irrespective of architecture. In summary, we cannot test all possible models, but have good grounds for proposing that feedforward CNNs cannot explain the flexible range of Bouma's window.

It is important to note that, contrary to our previous work [15, 16], we did not pick the stimuli to pit models against each other. The GA procedure produced the stimuli in a bottom-up fashion. As a limitation for pooling models, we cannot rule out that running the procedure for more generations may lead to "good" configurations that were not found using only 6 generations. However, there are principled reasons that explain why pooling models do not reproduce human results. Indeed, without grouping and segmentation to "rescue" the target from the flankers, all elements within Bouma's window would decrease performance in those models. Grouping and segmentation seem crucial to explain crowding in general [10, 15, 44, 48]. Moreover, it is known that texture models and other models based on pooling do not reproduce human grouping and segmentation [15, 16, 43, 52, 53]. Hence, it seems unlikely that simply adding generations in the GA algorithm could lead to human-like behaviour. Moreover, even if these models did find interesting configurations after a thousand generations, they would not reproduce an important behaviour, namely, rapid convergence of the GA.

How exactly grouping is implemented in humans is an open question. Here, we have used two different models that include a grouping mechanism. The grouping mechanism in the Laminart model is the formation of illusory contours between well-aligned edges that favour the parsing of visual elements into different layers of the network. This model works particularly well for the kind of displays that are used in Van der Burg et al. [35], because vertical and horizontal elements placed on a regular grid are either perfectly aligned or not aligned at all. However, this mechanism breaks down for more naturalistic stimuli, in which the complexity of low-level edges leads to an excess of illusory contours and, therefore, to bad segmentation. Capsule networks use a fundamentally different mechanism in which grouping is determined by recurrently maximizing the agreement between how neurons interpret a stimulus [45]. This mechanism is much more general than the Laminart model and is a promising candidate as a general framework to understand grouping and segmentation [43]. There are many more possibilities. For example, Linsley et al. [54] proposed another general recurrent grouping mechanism that is scalable to solve complex visual tasks at a state of the art level.

Future research will pit different models of grouping and segmentation against each other. (Un) crowding is one testbed in this respect, but there are many others, for example involving texture segmentation [52, 53], naturalistic image segmentation [54] or spatiotemporal grouping and segmentation [55]. Given the importance of grouping and segmentation, investigating which models can explain these results is an important step towards a better understanding of human vision.

Supporting information

S1 Appendix. Bouma's law model. Detailed description of the model, more details about the results.
(PDF)

S2 Appendix. Population coding model. Detailed description of the model.
(PDF)

S3 Appendix. Texture model. Detailed description of the model.
(PDF)

S4 Appendix. CNN classifier. Detailed description of the model.
(PDF)

S5 Appendix. Contour segmentation model ("Laminart"). Detailed description of the model.
(PDF)

S6 Appendix. Capsule network. Description of the model, including simulation of control models.
(PDF)

S7 Appendix. Two-stage model ("Popart"). Detailed description of the model.
(PDF)

S8 Appendix. Human experiment for proportion measure. Detailed description of the experiment.
(PDF)

S9 Appendix. Pooling model controls. Control simulations for all pooling models.
(PDF)

S10 Appendix. Quantitative similarity measurements. Assessment of the similarity between the model results and the human data, for Fig 3.
(PDF)

S11 Appendix. Fine-grained version of the selection measures. Detailed version of the rightmost column of Fig 3.
(PDF)

Author Contributions

Conceptualization: Alban Bornet, Adrien Doerig, Erik Van der Burg.

Formal analysis: Alban Bornet, Adrien Doerig.

Investigation: Alban Bornet.

Methodology: Alban Bornet, Adrien Doerig.

Software: Alban Bornet, Adrien Doerig, Erik Van der Burg.

Supervision: Michael H. Herzog, Gregory Francis, Erik Van der Burg.

Visualization: Alban Bornet.

Writing – original draft: Alban Bornet, Adrien Doerig.

Writing – review & editing: Alban Bornet, Adrien Doerig, Michael H. Herzog, Gregory Francis, Erik Van der Burg.

References

1. Gattass R, Gross CG, Sandell JH. Visual topography of V2 in the macaque. J Comp Neurol. 1981; 201 (4):519–39. https://doi.org/10.1002/cne.902010405 PMID: 7287933
2. Gattass R, Sousa AP, Gross CG. Visuotopic organization and extent of V3 and V4 of the macaque. J Neurosci. 1988; 8(6):1831–45. https://doi.org/10.1523/JNEUROSCI.08-06-01831.1988 PMID: 3385477
3. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962; 160(1):106. https://doi.org/10.1113/jphysiol.1962.sp006837 PMID: 14449617
4. Hubel DH, Wiesel TN. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. J Neurophysiol. 1965; 28(2):229–89.
5. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, et al. Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems. 1990. p. 396–404.
6. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999; 2 (11):1019–25. https://doi.org/10.1038/14819 PMID: 10526343
7. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell. 2007; 29(3):411–26. https://doi.org/10.1109/TPAMI.2007.56 PMID: 17224612

8. Ungerleider LG, Haxby JV. 'What'and 'where'in the human brain. Curr Opin Neurobiol. 1994; 4(2):157–65. https://doi.org/10.1016/0959-4388(94)90066-3 PMID: 8038571

9. Wallis G, Rolls ET. Invariant face and object recognition in the visual system. Prog Neurobiol. 1997; 51 (2):167–94. https://doi.org/10.1016/s0301-0082(96)00054-8 PMID: 9247963

10. Herzog MH, Sayim B, Chicherov V, Manassi M. Crowding, grouping, and object recognition: A matter of appearance. J Vis. 2015; 15(6):5–5. https://doi.org/10.1167/15.6.5 PMID: 26024452

11. Herzog MH, Thunell E, Ögmen H. Putting low-level vision into global context: Why vision cannot be reduced to basic circuits. Vision Res. 2016; 126:9–18. https://doi.org/10.1016/j.visres.2015.09.009 PMID: 26456069

12. Herzog MH, Clarke AM. Why vision is not both hierarchical and feedforward. Front Comput Neurosci. 2014; 8:135. https://doi.org/10.3389/fncom.2014.00135 PMID: 25374535

13. Saarela TP, Westheimer G, Herzog MH. The effect of spacing regularity on visual crowding. J Vis. 2010; 10(10):17–17. https://doi.org/10.1167/10.10.17 PMID: 20884482

14. Herzog MH, Manassi M. Uncorking the bottleneck of crowding: a fresh look at object recognition. Curr Opin Behav Sci. févr 2015; 1:86–93.

15. Doerig A, Bornet A, Rosenholtz R, Francis G, Clarke AM, Herzog MH. Beyond Bouma's window: How to explain global aspects of crowding? PLoS Comput Biol. 2019; 15(5):e1006580. https://doi.org/10.1371/journal.pcbi.1006580 PMID: 31075131

16. Doerig A, Bornet A, Choung OH, Herzog MH. Crowding reveals fundamental differences in local vs. global processing in humans and machines. Vision Res. 2020; 167:39–45. https://doi.org/10.1016/j.visres.2019.12.006 PMID: 31918074

17. Parkes L, Lund J, Angelucci A, Solomon JA, Morgan M. Compulsory averaging of crowded orientation signals in human vision. Nat Neurosci. 2001; 4(7):739–44. https://doi.org/10.1038/89532 PMID: 11426231

18. Pelli DG, Tillman KA. The uncrowded window of object recognition. Nat Neurosci. 2008; 11(10):1129–35. https://doi.org/10.1038/nn.2187 PMID: 18828191

19. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. J Vis. 2019; 19(7):15–15. https://doi.org/10.1167/19.7.15 PMID: 31348486

20. Wilson HR, Wilkinson F, Asaad W. Concentric orientation summation in human form vision. Vision Res. 1997; 37(17):2325–30. https://doi.org/10.1016/s0042-6989(97)00104-1 PMID: 9381668

21. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. Vision Res. 1973; 13(4):767–82. https://doi.org/10.1016/0042-6989(73)90041-2 PMID: 4706350

22. Levi DM. Crowding—An essential bottleneck for object recognition: A mini-review. Vision Res. 2008; 48 (5):635–54. https://doi.org/10.1016/j.visres.2007.12.009 PMID: 18226828

23. Pelli DG, Palomares M, Majaj NJ. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. J Vis. 2004; 4(12):12–12. https://doi.org/10.1167/4.12.12 PMID: 15669917

24. Strasburger H, Harvey LO, Rentschler I. Contrast thresholds for identification of numeric characters in direct and eccentric view. Percept Psychophys. 1991; 49(6):495–508. https://doi.org/10.3758/bf03212183 PMID: 1857623

25. Livne T, Sagi D. Configuration influence on crowding. J Vis. 2007; 7(2):4–4. https://doi.org/10.1167/7.2.4 PMID: 18217819

26. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. J Vis. 2012; 12(10):13–13. https://doi.org/10.1167/12.10.13 PMID: 23019118

27. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. J Vis. 2013; 13 (13):10–10. https://doi.org/10.1167/13.13.10 PMID: 24213598

28. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. J Vis. 2016; 16(3):35–35. https://doi.org/10.1167/16.3.35 PMID: 26913627

29. Poder E. Crowding, feature integration, and two kinds of "attention". J Vis. 2006; 6(2):7–7.

30. Saarela TP, Sayim B, Westheimer G, Herzog MH. Global stimulus configuration modulates crowding. J Vis. 2009; 9(2):5–5. https://doi.org/10.1167/9.2.5 PMID: 19271915

31. Saarela TP, Herzog MH. Time-course and surround modulation of contrast masking in human vision. J Vis. 2008; 8(3):23–23. https://doi.org/10.1167/8.3.23 PMID: 18484829

32. Manassi M, Whitney D. Multi-level crowding and the paradox of object recognition in clutter. Curr Biol. 2018; 28(3):R127–33. https://doi.org/10.1016/j.cub.2017.12.051 PMID: 29408262

33. Whitney D, Haberman J, Sweeny TD. 49 From Textures to Crowds: Multiple Levels of Summary Statistical Perception. 2014;

34. Whitney D, Levi DM. Visual crowding: A fundamental limit on conscious perception and object recognition. Trends Cogn Sci. 2011; 15(4):160–8. https://doi.org/10.1016/j.tics.2011.02.005 PMID: 21420894

35. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. J Exp Psychol Hum Percept Perform. 2017; 43(4):690. https://doi.org/10.1037/xhp0000337 PMID: 28182476

36. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press; 1992.

37. Van der Burg E, Cass J, Theeuwes J, Alais D. Evolving the stimulus to fit the brain: A genetic algorithm reveals the brain's feature priorities in visual search. J Vis. 2015; 15(2):8–8. https://doi.org/10.1167/15.2.8 PMID: 25761347

38. Kong G, Alais D, Van der Burg E. Competing distractors facilitate visual search in heterogeneous displays. PloS One. 2016; 11(8):e0160914. https://doi.org/10.1371/journal.pone.0160914 PMID: 27508298

39. Van de Weijgert M, Van der Burg E, Donk M. Attentional guidance varies with display density. Vision Res. 2019; 164:1–11. https://doi.org/10.1016/j.visres.2019.08.001 PMID: 31401217

40. Van den Berg R, Roerdink JB, Cornelissen FW. A neurophysiologically plausible population code model for feature integration explains visual crowding. PLoS Comput Biol. 2010; 6(1):e1000646. https://doi.org/10.1371/journal.pcbi.1000646 PMID: 20098499

41. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. J Vis. 2009; 9(12):13–13. https://doi.org/10.1167/9.12.13 PMID: 20053104

42. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–105.

43. Doerig A, Schmittwilken L, Sayim B, Manassi M, Herzog MH. Capsule networks as recurrent models of grouping and segmentation. PLOS Comput Biol. 2020; 16(7):e1008017. https://doi.org/10.1371/journal.pcbi.1008017 PMID: 32692780

44. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. Psychol Rev. 2017; 124(4):483. https://doi.org/10.1037/rev0000070 PMID: 28437128

45. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Advances in neural information processing systems. 2017. p. 3856–66.

46. Vickery TJ, Shim WM, Chakravarthi R, Jiang YV, Luedeman R. Supercrowding: Weakly masking a target expands the range of crowding. J Vis. 1 févr 2009; 9(2):12–12. https://doi.org/10.1167/9.2.12 PMID: 19271922

47. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989; 2(5):359–66.

48. Bornet A, Kaiser J, Kroner A, Falotico E, Ambrosano A, Cantero K, et al. Running large-sca fle simulations on the Neurorobotics Platform to understand vision-the case of visual crowding. Front Neurorobotics. 2019; 13:33.

49. Van der Burg E, Reynolds A, Cass J, Olivers C. Visual Crowding Does Not Scale With Eccentricity for Densely Cluttered Displays. In: PERCEPTION. SAGE PUBLICATIONS LTD 1 OLIVERS YARD, 55 CITY ROAD, LONDON EC1Y 1SP, ENGLAND; 2019. p. 27–27.

50. Reynolds JH, Heeger DJ. The normalization model of attention. Neuron. 2009; 61(2):168–85. https://doi.org/10.1016/j.neuron.2009.01.002 PMID: 19186161

51. Geirhos R, Meding K, Wichmann FA. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. ArXiv Prepr ArXiv200616736. 2020;

52. Herrera-Esposito D, Coen-Cagli R, Gomez-Sena L. Flexible contextual modulation of naturalistic texture perception in peripheral vision. bioRxiv. 2020;

53. Wallis TS, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M. Image content is more important than Bouma's Law for scene metamers. ELife. 2019; 8:e42512. https://doi.org/10.7554/eLife.42512 PMID: 31038458

54. Linsley D, Kim J, Serre T. Sample-efficient image segmentation through recurrence. ArXiv Prepr ArXiv181111356. 2018;

55. Drissi-Daoudi L, Doerig A, Herzog MH. Feature integration within discrete time windows. Nat Commun. 2019; 10(1):1–8. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773

S1 Appendix: Bouma's law model

To set a basis for our analysis, we used a model that assumes Bouma's law (1) holds true in dense displays. In this model (Fig A, top), any flanker in the dense display creates the same amount of interference as it would do in a sparse display. To set interaction weights between the flankers and the target, we used the data from the sparse display experiment of Van der Burg et al. (1; Fig 2a in main text, bottom). Based on this data, we defined interaction weights for any flanker as the performance drop that it would cause in the sparse display experiment, and the total interaction T as the sum of the weights of all flankers in the display. For each display, we defined the probability for the model to make a correct response as in Eq. 1.

$$P_{correct} = max\left[P_{unflanked} \cdot (1 - A \cdot T), 0.5\right] \quad (1)$$

$P_{unflanked}$ comes from the sparse display experiment in Van der Burg et al. (2) and is the average proportion of correct responses without flankers and A is a global gain for the interaction weights. A was set to 1.0 for sparse displays but was lowered to 0.3 for dense displays to avoid the model being always at chance level. It was tuned to obtain approximately 67% performance for the first generation in the GA procedure. Performance for each display was defined as the probability of correct responses.

Note that this model was used in Van der Burg et al. (2), to investigate whether the GA procedure was able to produce behaviour consistent with Bouma's law in the first place. However, directly using the probability of correct responses to select the best displays at each generation, without simulating trials, might have discounted variability in the evolution process of the GA. Hence, for completeness, we ran a second version of the model that, instead, selected the best displays based on the simulation of 12 trials (still using the probability of correct responses as in Eq. 1, the first version of the model corresponds to running the second one with an infinite number of trials).

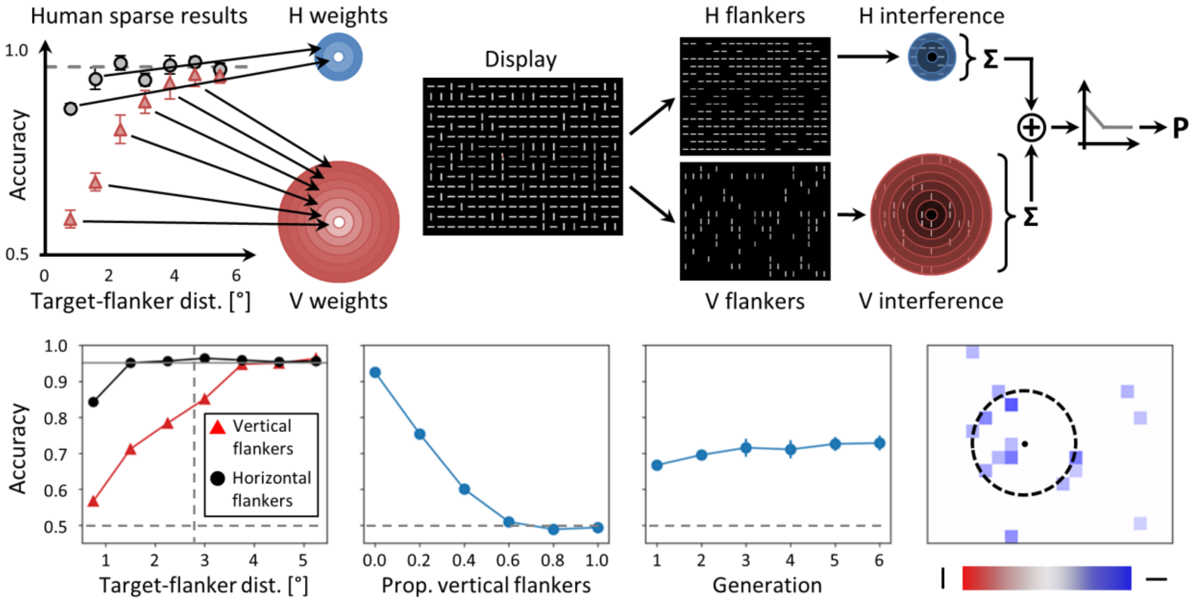

Fig A. Top. Bouma model. Flanker-target distance-dependent weights were defined as how much performance dropped from the unflanked level in the sparse display experiment of Van der Burg et al. (2). The input of the model is an array of 15 by 19 bits that encodes each flanker orientation. Spatial units in the model are implicitly encoded by setting the weights of each flanker location to the human data. For each display, the probability of correct response is a decreasing function of the sum of its flankers' weights (see Eq. 1). **Bottom.** Results obtained with the second version of the Bouma model (same description as in Fig 3 in the main text).

The results for both ways of selecting the best displays between generations are shown in Fig 3 in the main text (2nd row) for the first version and in Fig A (bottom) for the second version. Both versions reproduced human results for the sparse display and the proportion measures. Model performance improved as much as in the human experiment during the GA procedure for the first version, but the second version produced only a minor improvement. This may be due to the variability added by the selection process in the second version of the model. In consequence, the GA procedure did not highlight any specific location in the selection measure for the second version of the model, whereas essentially all elements inside Bouma's window were highlighted for the first version. In summary, both versions of the model did not account for the shrinking of Bouma’s window.

References

1. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. Vision Res. 1973;13(4):767-82.

2. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. J Exp Psychol Hum Percept Perform. 2017;43(4):690.

S2 Appendix: Population coding model

The population coding model (1) provides a physiologically plausible description of the spatial integration of orientation signals and accounts for various aspects of visual crowding. In this model, a population of orientation-sensitive neurons encodes the content of each location in the stimulus array (Fig A). These neuron populations constitute the first layer of the model. Neurons in the second layer pool stimulus information locally, using a weighted summation of the population activities in the first layer. The weighting fields are expressed in cortical coordinates and hence depend on the population eccentricity. Then, orientation is decoded from the activity of the population in the second layer that corresponds to the target location. A mixture of von Mises distribution is fit to the population activity and the maximum value of the fitted function is taken as the decoded orientation. For each display, performance was computed as the proportion of decoded orientations of same sign as the target orientation.

Model parameters were the same as in (1), except for the pooling range that was adapted to produce Bouma's law in sparse displays. For dense displays, the model was very close to chance level, because the pooled activity from horizontal flankers was so large that it overwhelmed the activity coming directly from the target. To solve this issue, we added a prior to select target orientation: the value of the fitted von Mises mixture function was set to zero for any orientation outside the range [-45°, 45°], before it was used to decode the target orientation. However, even using the former prior, simply because there were too many flankers that were pooled in dense displays, the model was too close to chance level for the GA to work (performance could not increase during the GA procedure). To help the model reaching 67% of accuracy in the first generation, we increased the target orientation to ±10° (instead of ±5°), for dense displays only.

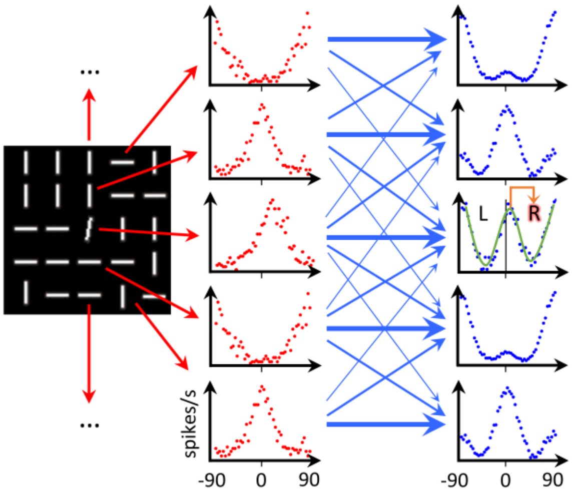

Fig A. Population coding model. Populations of orientation-selective neurons encode the content of every element in the display array (red arrows). The input of the model is an array of 15 by 19 bits that encodes each flanker orientation (only a subset is shown here). Then, the activity responsible for each location of the array is pooled to a second layer of neuron populations (blue arrows). Pooling weights (represented here by the thickness of the arrows) depend on the cortical distances between the populations. Am important parameter of the model is the cortical pooling range (which defines spatial units in the model) and was fitted to yield Bouma's law in sparse displays. Finally, the target orientation is decoded from the second layer activity by fitting a mixture of von Mises distributions (green) to the activity of the population responsible for the target. The sign of the target orientation is used to report a left or a right target.

Results obtained with the model are shown in Fig 3 in the main text (3rd row). The model reproduced human behaviour very well for the sparse display measure. For the proportion measure, the model performed better than humans for small proportions of vertical flankers. This may have been due to the prior that we added to the decoding process of the model. The GA procedure increased model performance dramatically, even to a larger extent than in the human experiment. The selection measure highlighted a large portion of the locations inside Bouma's window, which is not in accordance with the human results. Note that there was an inward-outward anisotropy (2) in the highlighted locations, i.e., flankers on the peripheral side of the target had more impact than flankers on the foveal side. This can be explained, because the model takes cortical magnification into account: pooling distances are expressed in cortical units and hence, pooling has a larger range for populations located in the periphery than near the fovea. In summary, this model reproduces human results for all measures, except the selection measure.

References

1. Van den Berg R, Roerdink JB, Cornelissen FW. A neurophysiologically plausible population code model for feature integration explains visual crowding. PLoS Comput Biol. 2010;6(1):e1000646.

2. Toet A, Levi DM. The two-dimensional shape of spatial interaction zones in the parafovea. Vision Res. 1992;32(7):1349-57.

S3 Appendix: Texture model

Texture models (1) iteratively update an array of pure noise, until an image is produced that matches a specific set of statistics computed from the model's visual input. These models are seen as models of vision, because they provide a very efficient way to encode visual information in the brain, even for natural images (which are rather complex in terms of visual content, like our dense displays). Balas et al. (2) proposed that crowding is the result of such statistics being computed over pooling regions. They proposed to use the model of Portilla et Simoncelli (1) over a Bouma-sized patch centred on the target to generate textures whose content reflect the amount of crowding associated to the flanker pattern present in the input image. Rosenholtz et al. (3) proposed to improve this model by computing the statistics over many tiled regions whose size grow with eccentricity. However, we did not use the latter model because it was computationally too heavy: given the stimulus dimensions, it would have taken approximately 2 years to run the GA procedure on our lab computer.

We used the code available at https://github.com/LabForComputationalVision/textureSynth to produce the same kind of Bouma-sized textures, using the displays of Van der Burg et al. (4). For each display trial, we generated a texture and decoded whether the target was oriented to the left or to the right, using a template match algorithm (Fig Aa). The algorithm uses left and right target templates and looks for the best match over the whole texture. Every trial produced a different texture image, because the generative process is stochastic. The performance of the model was then the fraction of correct responses over the trials.

The reason why an algorithm was used instead of human observers looking at the textures, (as in 40) is that, to create new generations of displays, the GA procedure must know the performance associated to the parent displays. Hence, it would have required the textures to be generated during the experiment, which would have added about 1 minute of texture computation between every button press in a human experiment, making it last about 64 hours per human participant. To make sure that our template match algorithm captured human performance qualitatively, we ran an experiment in which humans looked freely at the Bouma-sized textures generated by the model for dense displays in which the proportion of vertical flankers was varied. The task was to decide whether the texture came from a display that contained a target tilted to the left or to the right compared to vertical. We fitted the parameters of the template match algorithm to match human performance (Fig Ab).

As with the population coding model, the results of the texture model for dense displays were too close to chance level. Therefore, we increased the target orientation to ±15° (instead of ±5°), for dense displays only, so that performance was around 67% for the first generation of displays in the GA procedure. Note that this was not the case with the validation experiment we ran to produce the panel in Fig Ab.

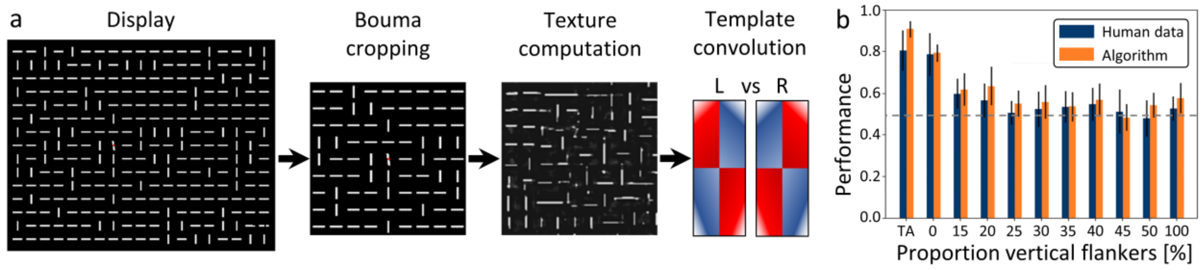

Fig A. a. Texture model. The input to the model is an actual image of the visual stimulus. Spatial units in the model are defined by the resolution of the stimulus, which was set to 30 pixels per degree. First, the stimulus display is cropped, so that only a Bouma-sized patch around the target is sent to the texture model. Then, the model iteratively matches a set of statistics between the input patch and the output texture. Finally, an algorithm chooses whether the texture comes from a display in which the target is tilted to the left or to the right by convolving left and right filters to the output texture and looking for the maximal match. **b.** Comparison between the template match algorithm and experimental results in which human observers discriminated the target orientation from the output textures in free-viewing conditions, for different proportions of vertical flankers (TA stands for target alone). The algorithm captures human behaviour.

The results obtained with the model are shown in Fig 3 in the main text (4th row). For the sparse display measure, the model performance did not show a clear dependence on target-flanker distance, aside from the performance bump that happened when the flankers went outside the cropping range. This suggests that interference in this model does not depend on the relative location of elements, which is in contradiction with human results. This was already a hint that the model would not highlight special configuration in the GA procedure but would at best behave like the second version of the Bouma model. As expected, although the model reproduced human results for the proportion measure, performance did not improve in the GA procedure and the selection measure did not highlight any location, exactly as with the second version of the Bouma model (Fig A in S1 Appendix, bottom). In summary, the texture model only reproduced human results for the proportion measure.

References

1. Portilla J, Simoncelli EP. A parametric texture model based on joint statistics of complex wavelet coefficients. Int J Comput Vis. 2000;40(1):49-70.

2. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. J Vis. 2009;9(12):13-13.

3. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. J Vis. 2019;19(7):15-15.

4. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. J Exp Psychol Hum Percept Perform. 2017;43(4):690.

S4 Appendix: CNN classifier

Deep feedforward convolutional neural networks (CNNs) share many similarities with humans in their architecture, in their activity patterns (1,2), as well as in the performance they reach in a large number of visual tasks (3,4). Here, we used the same method as in (5), testing AlexNet (6) as a representative of CNNs, because it is often used as a model of the human visual system (7–10). The weights of AlexNet were already trained on ImageNet (11). To perform the crowding task, we trained different classifiers to decode target orientation (left or right) based on the activity of each layer of the network. The training set was made of images that contained both the target and an array of vertical and horizontal flankers (Fig A). Only the weights of the classifier were affected by the training phase. In the image samples of the training set, the target never overlapped with the flanker array. After this training phase, the model used in the GA procedure consisted in AlexNet, plus the classifier whose layer gave the best fit of Bouma's law for sparse displays (which was the fourth layer). The performance of the model was then simply the fraction of correct classifications over the trials.

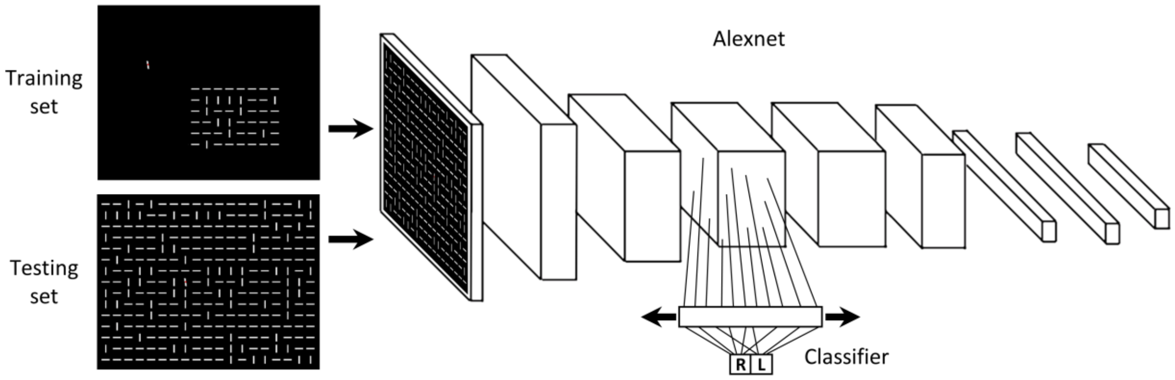

Fig A. CNN classifier. The input to the model is an actual image of the visual stimulus. Spatial units in the model are defined by the resolution of the stimulus, which was set to 15 pixels per degree, i.e., the maximum value given the input shape that AlexNet accepts (227 by 227 pixels). The stimulus display is processed by the architecture of Alexnet. On top of each layer, a decoder was trained to discriminate between left or right targets from the layer activity. The weights of Alexnet (which have been previously trained on ImageNet) did not change during the training process. The training set was composed of samples containing the target alone and an array of vertical and horizontal flankers that never overlapped with the target. The loss function of the classifier was the cross-entropy on target classification. After training the classifiers, the whole model was tested with the four measures described in the Methods section. The reported results came from the trained classifier put on top of the layer that gave the best fit of Bouma's law in the sparse display measure. Adapted with permission from (5).

The results obtained with the model are shown in Fig 3 in the main text (5th row). None of the layers reproduced Bouma's law qualitatively in the sparse display measure. We report all measures that we obtained with a classifier put on top of the fourth layer of Alexnet, which gave the least bad fit, and whose receptive field size roughly matches the size of Bouma's window. For dense displays, the model performance generally decreased with the proportion of vertical flankers. However, the model was at chance level with 100% of horizontal flankers. During the GA procedure, model performance increased only marginally. The selection measure did not highlight any specific location that was crucial for this improvement. In summary, the CNN classifier replicated none of the human results.

References

1. Lindsey J, Ocko SA, Ganguli S, Deny S. A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. ArXiv Prepr ArXiv190100945. 2019;

2. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818-33.

3. Eslami SA, Rezende DJ, Besse F, Viola F, Morcos AS, Garnelo M, et al. Neural scene representation and rendering. Science. 2018;360(6394):1204-10.

4. Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K. Detectron. 2018.

5. Doerig A, Bornet A, Choung OH, Herzog MH. Crowding reveals fundamental differences in local vs. global processing in humans and machines. Vision Res. 2020;167:39-45.

6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097-105.

7. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014;10(11):e1003915.

8. Kietzmann TC, McClure P, Kriegeskorte N. Deep neural networks in computational neuroscience. BioRxiv. 2018;133504.

9. VanRullen R. Perception science in the age of deep neural networks. Front Psychol. 2017;8:142.

10. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci. 2014;111(23):8619-24.

11. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248-55.

S5 Appendix: Contour segmentation model (“Laminart”)

The Laminart model (1) is a spiking neural network that computes illusory contours between aligned edges. In the model, grouping is crucial. Elements linked by any contour (illusory or real) are grouped together by dedicated neural populations. Stimuli are first segmented into different groups by the network's dynamics and, subsequently, elements within a group interfere (Fig A). Importantly, crowding is weak when the target belongs to a different group than most flankers, and strong otherwise. The segmentation process is triggered by local selection signals whose activity then spreads along connected contours. The location of the selection signals determines the output of the segmentation process.

It is very time consuming to run the model (for our displays, it would need to simulate several millions of spiking neurons for each display trial) and cannot go through the whole GA procedure in a realistic amount of time. However, exploiting the fact that the flankers are exclusively vertical or horizontal, we built a faster segmentation algorithm that reproduce the model behaviour for the displays used in Van der Burg et al. (2). For each display, the algorithm links neighbouring bars whenever they or their tips are aligned (Fig A, right). The different groups are defined as all disconnected sets of bars that are linked by the former procedure. This corresponds exactly to the behaviour of the full Laminart model but requires much less time to run.

At each trial, the algorithm sends selection signals that segment any group that is reached. In Francis et al. (1), because the visual stimuli tested with the model consisted of a vernier target flanked on both sides, two selection signals were sent at each trial, one on each side of the target. Here, because the flankers lie on all sides of the target, four selection signals are sent around the target at each trial. The segmentation layer that contains the target was used to compute target-flanker interference. For each trial, the total interference, T, was defined exactly as in the Bouma model, and a choice was made about the orientation of the target, with a probability of correct response defined by Equation 1 in S1 Appendix. The only difference with the Bouma model is that, thanks to the segmentation process, a single gain A, was used for sparse and dense displays, without preventing the GA procedure to work. The performance for each display was defined as the fraction of correct responses over the trials.

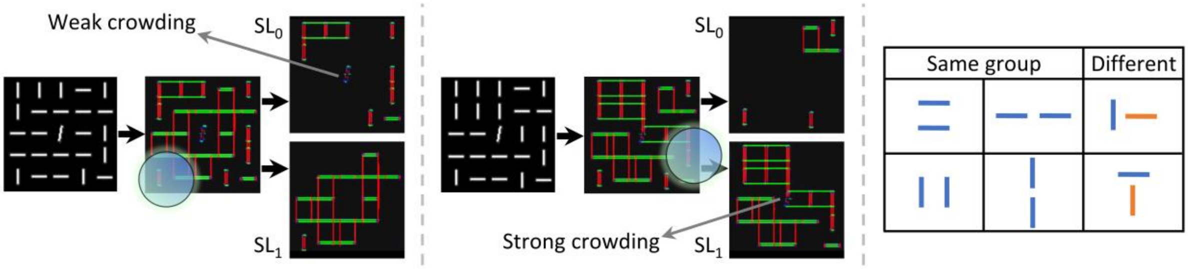

Fig A. Laminart model. In the original model (1), the stimulus is processed by an array of orientation-selective feature detectors. Coloured pixels that are depicted in the images correspond to the most active oriented cell at that location (red, green, blue, purple and turquoise for vertical, horizontal, oblique, almost vertical and almost horizontal orientations). Recurrent connections compute illusory contours between well-aligned edges. Elements that are linked by illusory contours belong to the same group. Then, local, top-down selection signals (blue circles) trigger a recurrent segmentation process, parsing the visual input in different segmentation layers (here, SL_0 and SL_1, but there can be more segmentation layers). After dynamic processing, all elements that are linked, through an actual or an illusory contour, to a location that is touched by a selection signal are parsed to the corresponding segmentation layer. Crowding is computed simply by applying the Bouma model to the segmentation layer that contains the target. **Left.** If only a few flankers are segmented with the target, crowding is weak. **Center.** If the target is linked with a large group of flankers through illusory contours, crowding is strong. **Right.** Because it would have taken too long to simulate the model for large displays, the segmentation process was replaced by an algorithm that reproduces its behaviour, given the simplicity of the stimuli involved in Van der Burg et al. (2). The algorithm assigns all elements to groups by linking pairs of well-aligned edges. After sending a selection signal, all groups of elements that are reached appear in the corresponding segmentation layer. The input of the grouping and segmentation algorithm is an array of 15 by 19 bits that encodes each flanker orientation, but the initial selection signal process is simulated using an actual image of the stimulus. Spatial units in the model are defined by the size of the selection signals, which have a radius of 1 degree in the original version of the Laminart model. The resolution of the image, set to 15 pixels per degrees, is not crucial to the model but is used to determine whether or not a selection signal that partially overlaps with a flanker location would hit the flanker.

Results obtained with the Laminart model are shown in Fig 3 in the main text (6th row). The model reproduced Bouma's law simply because target-flanker interference was defined as in the Bouma model. The model reproduced human results for the proportion measure. During the GA procedure, performance increased with the generations. The selection measure revealed that the flanker locations that were crucial for this improvement were the target's nearest neighbours. This can be explained by the fact that, whenever all these crucial locations contain horizontal flankers, a "grouping shield" is created around the target (such as in Fig Aa, left), so that: a) no illusory contour can ever group flankers with the target; b) a segmentation signal has a large probability to hit a flanker that is linked to this shield, parsing many flankers to a different segmentation layer than the one of the target. For these reasons, the target's nearest neighbours were more crucial to determine crowding strength than in other models. In summary, this model replicated all human results well, but interference in the model was directly fitted to the sparse display data instead of proposing a mechanism.

References

1. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. Psychol Rev. 2017;124(4):483.
2. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. J Exp Psychol Hum Percept Perform. 2017;43(4):690.

S6 Appendix: Capsule network

Capsule networks are deep neural networks in which layers of neurons communicate through a recurrent process that implements grouping (Fig A). Each layer is made of many capsules, groups of neurons encoding specific features within their pattern of activity. Layers communicate through a time-consuming recurrent process called “routing by agreement” (1), in which each capsule in the lower layer predicts the activity of each capsule in the next layer. Grouping happens when many capsules agree that a certain higher-level capsule should be highly active: the corresponding higher-level capsule is activated and other higher-level capsules for which there is no agreement are shut down (Fig A, right). The entire network is trained end to end through backpropagation. Doerig et al. (2) showed that Capsule networks can explain uncrowding based on their grouping capabilities.

We trained the model for the GA procedure using a similar approach as in Doerig et al. (2). The Capsule network was first trained to recognize targets and groups of horizontal or vertical elements using a training set consisting of images that either contained a target in isolation or a rectangular array of 1 to 49 uniformly horizontal or vertical flankers. During the training phase, the Capsule network was also trained to discriminate between left and right targets (Fig A, left). The model was trained until it was able to classify the target with 67% of accuracy on a validation set composed of dense display arrays with 30% of vertical flankers. Note that only one of the 10 models we trained reached this performance level. After the training phase, this model was tested with sparse and dense displays. The performance was defined as the fraction of correct classifications over the trials. Note that only Bouma-sized crops were sent to the Capsule network during training, validation and testing. This was done for a better convergence of the training loss and because the training process would have required too much memory to fit on our computer with full stimulus arrays.

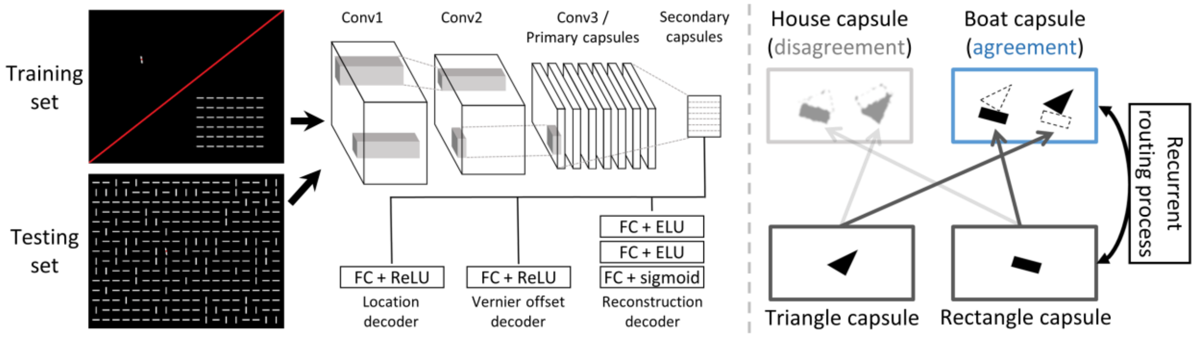

Fig A. Left. Capsule network. The input to the model is an actual image of the visual stimulus. Spatial units in the model are defined by the resolution of the stimulus, which was set to 15 pixels per degree (for computational reasons). The stimulus display is processed by a set of convolutional layers, conveying information to the primary capsules, which then projects to the secondary capsules. Routing by agreement happens between the primary and secondary capsules. In this case primary capsules encode visual elements (target, horizontal, vertical element) and the secondary capsules encode groups of visual elements. The output of the secondary capsules is sent to 3 different simple decoders (for stimulus reconstruction, stimulus location decoding and target orientation discrimination). The training set is composed of samples containing either the target alone or an array of exclusively vertical or horizontal flankers. The loss of the classifier is a combination of a reconstruction loss, and of cross-entropies on target classification and on stimulus location. In addition, a margin loss makes sure that the activity in the secondary capsules corresponds to the correct types of visual elements (target, horizontal, vertical group). After training the whole network end to end, we tested it with the four measures described in the Methods section, using the target orientation decoder to generate responses for each stimulus. **Right.** Routing by agreement. In this example, capsules in the lower layer encode basic shapes, and capsules in the higher layers encode objects. The activity pattern of each capsule encodes the characteristics of the input it is responsible for (size, location, orientation, etc.). Both primary capsule's outputs try to predict how activity is going to look in the secondary capsules. Because their predictions match in the boat capsule (dashed shapes vs. full shapes), the projection that lead to this agreement (dark arrows) is strengthened over time by the recurrent routing process. Because these same primary capsules do not agree with each other in the house capsule, this projection (light arrows) is weakened by the routing process. Adapted with permission from (2).

Results obtained with the Capsule network are shown in Fig 3 in the main text (7th row). Surprisingly, the model reproduced Bouma’s law qualitatively simply by being trained at identifying targets and flankers (albeit unflanked performance is higher than in humans). The model reproduced human results for the proportion measure as well. The GA procedure improved the performance of the Capsule network along the generations, and the selection measure showed that the flanker locations that were crucial for this improvement were just above and below the target. In summary, this model replicated all human results well, except that only the flankers directly above and below the target (and not those to the left and right) are highlighted by the selection measure. One caveat is that only one out of the 10 models we trained reached good target discrimination in dense displays.

To control for the importance of segmentation processes in Capsule networks, we added simulations of control versions of the model (see Fig. B), as what was done in Doerig et al. (2). Importantly, these control versions contain the same number of parameters as in the capsule network but do not instantiate any grouping process. The first control version is a Capsule network in which the capsule layers are replaced by a fully connected feedforward layer, yielding a standard feedforward CNN with three convolutional layers and a fully connected layer. The second version is the same network as in the first control version, but with added lateral recurrent connections in the fully connected layer of the feedforward CNN, yielding a network with three convolutional layers followed by a fully connected recurrent layer. The third version is the same network as in the first control version, but with added top-down recurrent connections feeding back from the final fully connected layer of the feedforward CNN to the layer below, yielding a network with three convolutional layers followed by a fully connected layer that feed back into the previous one. The control models did not reproduce the human data, highlighting the importance of grouping processes in the Capsule network.

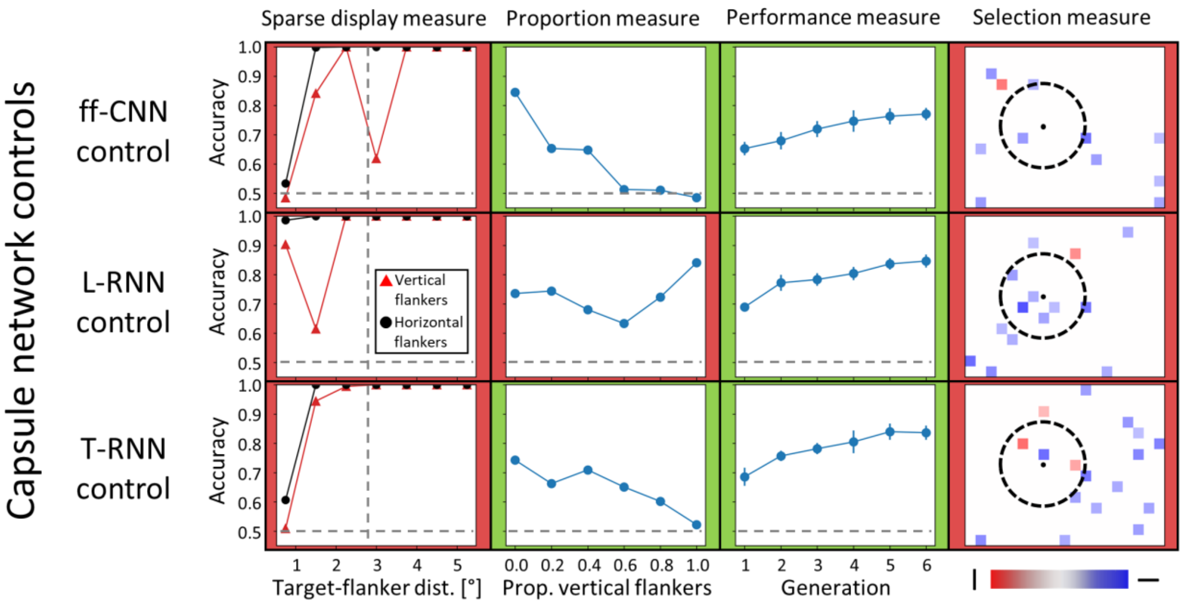

Fig B. Top. Results obtained with the feedforward CNN control version of the Capsule network model. **Center.** Results obtained with the lateral RNN control version of the Capsule network model. **Bottom.** Results obtained with the top-down RNN control version of the Capsule network model. The performance of all control models increases with the generations. However, all control models fail to reproduce the selection measure, as the GA procedure does not highlight any particular flanker location responsible for the performance improvements.

References

1. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Advances in neural information processing systems. 2017. p. 3856-66.

2. Doerig A, Schmittwilken L, Sayim B, Manassi M, Herzog MH. Capsule networks as recurrent models of grouping and segmentation. PLOS Comput Biol. 2020;16(7):e1008017.

S7 Appendix: Two-stage model ("Popart")

The contour segmentation model (Laminart model; 1) explained the configuration effects in dense displays very well, but the way to measure target-flanker interaction was simply to fit the experimental data of Van der Burg et al. (2) for sparse displays. On the other hand, the population coding model (3) naturally accounts for Bouma's law in sparse displays but does not replicate the preference measure in dense displays. For these reasons, we combined both models into a two-stage model (Fig A).

In this combination, the segmentation model acts as a grouping stage and selects *which* elements in the visual field are going to interfere with each other. Only the flankers that were parsed in the same group as the target are sent to the interference stage. The population coding model acts as an interference stage and determines *how* the elements that were selected during the first grouping stage interfere. The parameters of both models were kept the same as in their respective descriptions above. The only difference was that the performance measure of the segmentation model was now computed by feeding the content of the target's segmentation layer to the population coding model.

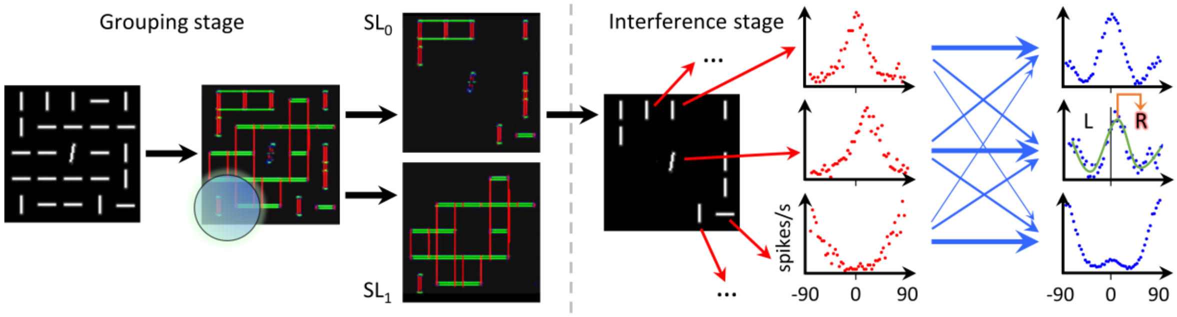

Fig A. Popart model. The model is composed of two stages. **Left.** Grouping stage. The Laminart model algorithm is used to parse the stimulus in different segmentation layers. **Right.** Interference stage. From the output of the segmentation algorithm, a new stimulus is built. Only the elements present in the segmentation layer that contains the target are processed by the population coding model to generate a response.

Results obtained with the model are shown in Fig 3 in the main text (last row). Thanks to the combination of both segmentation and population coding models, the Popart model qualitatively reproduces human results for all measures.

References

1. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. Psychol Rev. 2017;124(4):483.

2. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. J Exp Psychol Hum Percept Perform. 2017;43(4):690.

3. Van den Berg R, Roerdink JB, Cornelissen FW. A neurophysiologically plausible population code model for feature integration explains visual crowding. PLoS Comput Biol. 2010;6(1):e1000646.

S8 Appendix: Human experiment for proportion measure

We asked human participants to sit at 62 cm from an LCD screen (120 Hz refresh-rate), in a dimly lit room. The experiment was programmed and run using OpenSesame (1). The task was to discriminate between a left or a right target (tilted by 5°) presented in the periphery of the visual field. At each trial, as was done in Van der Burg et al. (2), the location of the target was either on the left or on the right of a central white fixation dot (0.1° radius). The possible target locations were indicated by two red dots (0.02° radius). These dots were visible during the whole experiment. When displayed, the target was embedded in an array of 15 rows by 19 columns of vertical or horizontal flankers (dense display, see Fig 2b in the main text). The target was always displayed at 6° of eccentricity (8th row, 8th column in the flanker array). A trial consisted of 500 ms during which the white and the red dots were presented alone, followed by 150 ms in which the target and the flanker array appeared, followed by an unlimited amount of time in which the observers could give their response by pressing a key. After the response was recorded, a new trial was initiated. The experiment consisted of 11 blocks (1 for practice) of 24 trials each. In each block, trials for each condition (0%, 20%, 40%, 60%, 80% or 100% of vertical elements in the flanker array) were mixed and evenly distributed (i.e., 6 trials per condition). At the end of each block, feedback was given to the observer as the proportion of correct responses in the performed block. We ran 7 participants in total, but we discarded 1 participant who was at chance level for all conditions. Results are shown in Fig 3 in the main text (top row, 2nd column). Participants gave oral consent before the experiment, which was conducted in accordance with the Declaration of Helsinki except for the preregistration (World Medical Organization, 2013) and was approved by the local ethics committee (Commission d'éthique du Canton de Vaud, protocol number: 164/14, title: Aspects fondamentaux de la reconnaissance des objets protocole général).

References

1. Mathôt S, Schreij D, Theeuwes J. OpenSesame: An open-source, graphical experiment builder for the social sciences. Behavior research methods. 2012;44(2):314-24.

2. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. Journal of Experimental Psychology: Human Perception and Performance. 2017;43(4):690.

S9 Appendix: Pooling model controls

In Fig 3 (main text), we tuned the model parameters to reproduce Bouma's law in sparse displays and showed that only the grouping models can shrink Bouma's window to the nearest neighbour distance in dense displays. Here, we included control simulations in which we instead tuned the parameters of the pooling models to reproduce human behaviour in dense displays (selection measure) and then examined their behaviour using sparse displays (Fig A). Each model's pooling range was decreased to the nearest neighbour distance by tuning the relevant parameters. For the Bouma model and the Popcode model (see S1 and S2 Appendices for more details), the pooling range is simply a scalar parameter that we modified. For the texture model (see S3 Appendix for more details) in which Bouma-sized crops of the stimuli were used, we simply used smaller crops. For the CNN classifier (see S4 Appendix for more details), we trained a classifier to decode left and right targets from the activity of the second layer of the CNN, whose receptive field size matches the nearest neighbour distance.

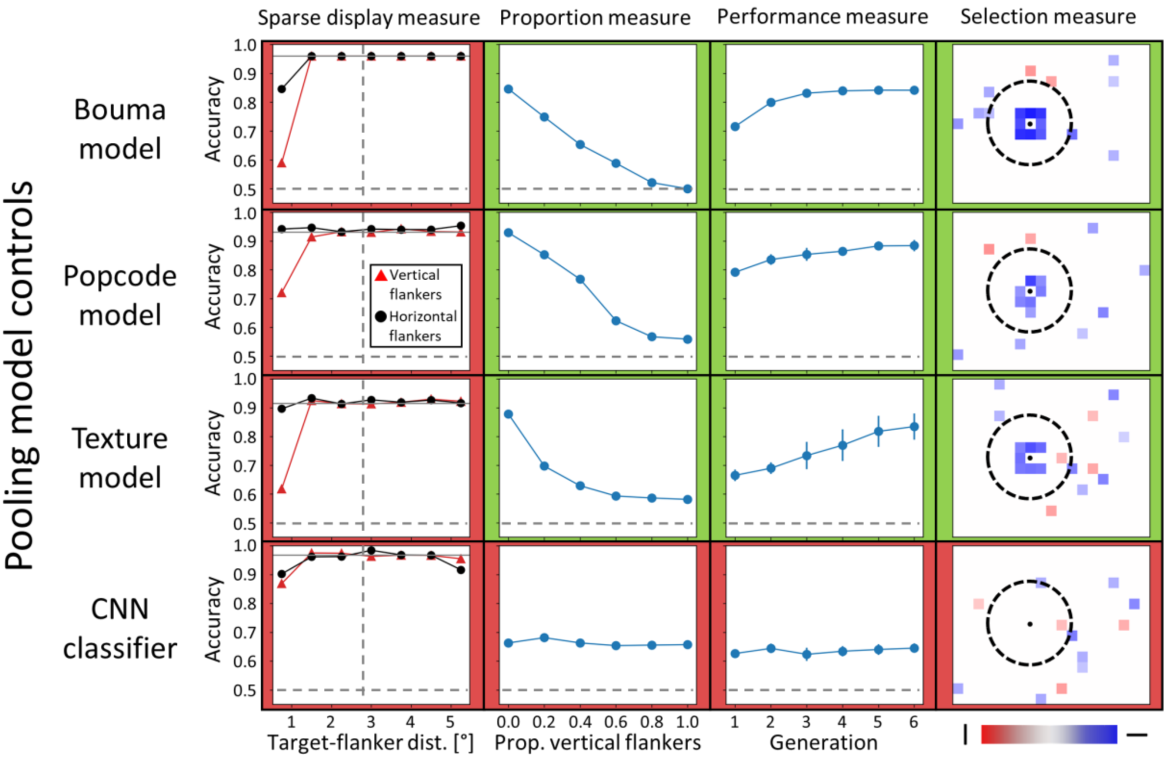

Fig A. Results obtained with the control simulations of the pooling models, in which the pooling range was adapted to reproduce human behaviour in the selection measure instead of the sparse display measure. The description of each measure is the same as in Fig 3 (main text). Except for the CNN classifier, in which the GA procedure did not improve performance, it was possible to adapt the interference range of the pooling models to reproduce human behaviour in the selection measure (the GA procedure improved performance, highlighting mostly the target nearest neighbours). However, the interference range was also shrunk in the sparse display measure (in sparse displays, performance was impaired only when the flankers were very close to the target).

The results in Fig A show that tuning pooling models to shrink Bouma's window in dense displays prevents them to reproduce Bouma's law in sparse displays. Indeed, the range of interference in sparse displays is also shrunk to the nearest neighbour distance, in contrast to the models that include a grouping component (see Fig 3). These results provide further evidence that a grouping stage is necessary to exhibit a small range of interference in dense display, while keeping a Bouma-sized range in sparse displays.

S10 Appendix: Quantitative similarity measurements

In Fig 3 (see main text), we compared how the different models we tested reproduce human behaviour in sparse and dense displays. To compare the different models, we qualitatively assessed the similarity between each model's measures and the corresponding human data. In Fig A, for completeness, we include *quantitative* assessments of the similarity between model and human behaviour. The similarity between the sparse display, proportion and performance measures and the corresponding human data was computed as the Pearson correlation coefficient between model and human performance. For the selection measure, we used the structural similarity index (1) because the output of this measure is a 2-dimensional image.

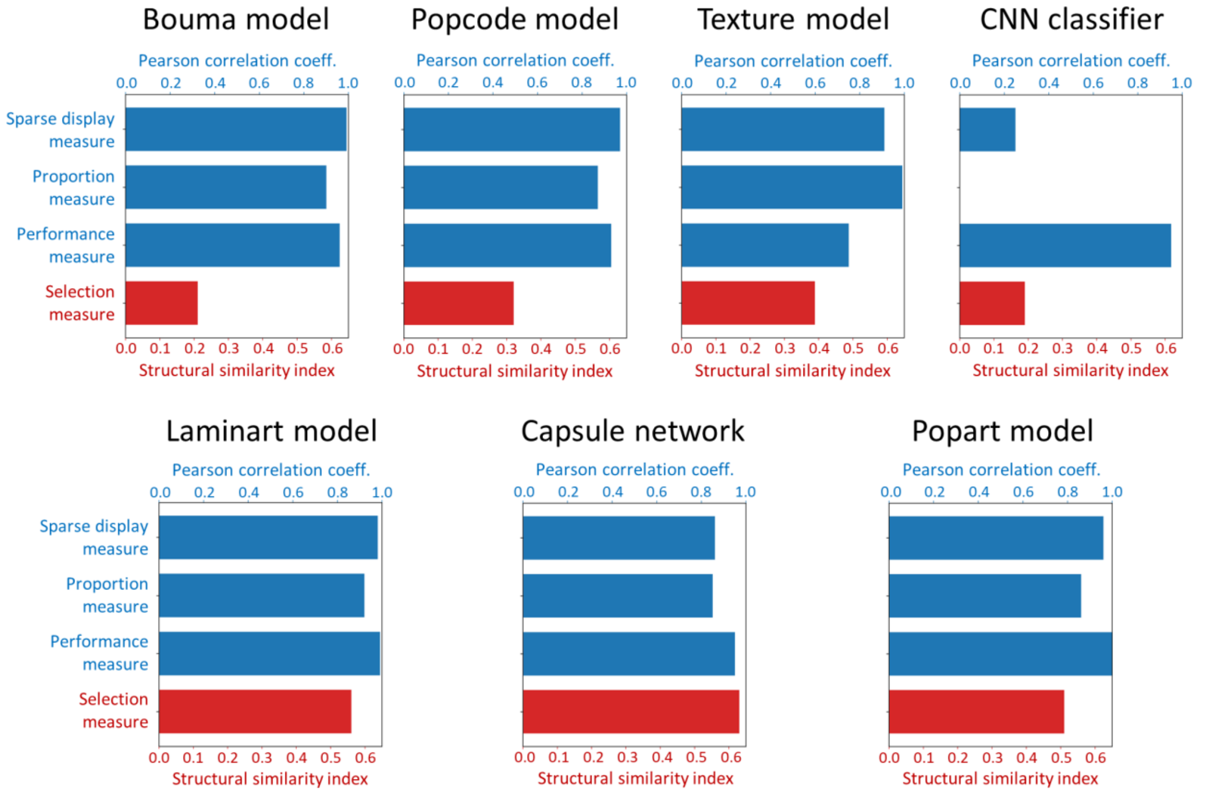

Fig A. Quantitative assessment of the similarity between the model results and the human data, for the different measures presented in Fig 3 (see main text). As described qualitatively in Fig 3, all models (except for the CNN classifier) reproduce human behaviour for the sparse display, proportion, and performance measures, but only the grouping models are able to reproduce human behaviour in the selection measure as well. Here, this claim is confirmed by our quantitative assessment method. Overall, the models obtain high and comparable similarly scores for the sparse display, proportion, and performance measures (except for the CNN classifier). However, the grouping models overall obtain higher similarity scores than pooling models for the selection measure.

References

1. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13(4):600-12.

S11 Appendix: Fine-grained version of the selection measures

In Fig 3 (see main text), the results of the selection measures are presented by cutting the values for which neither vertical nor horizontal flankers were significantly overrepresented by the GA procedure, compared to a random selection process. This allowed to highlight the range of interaction between the target and the flankers in dense displays. For completeness, we include a "fine-grained" version of these measures (for the human data as well as for the model results), in which the values are not put down to zero if not statistically significant (Fig A).

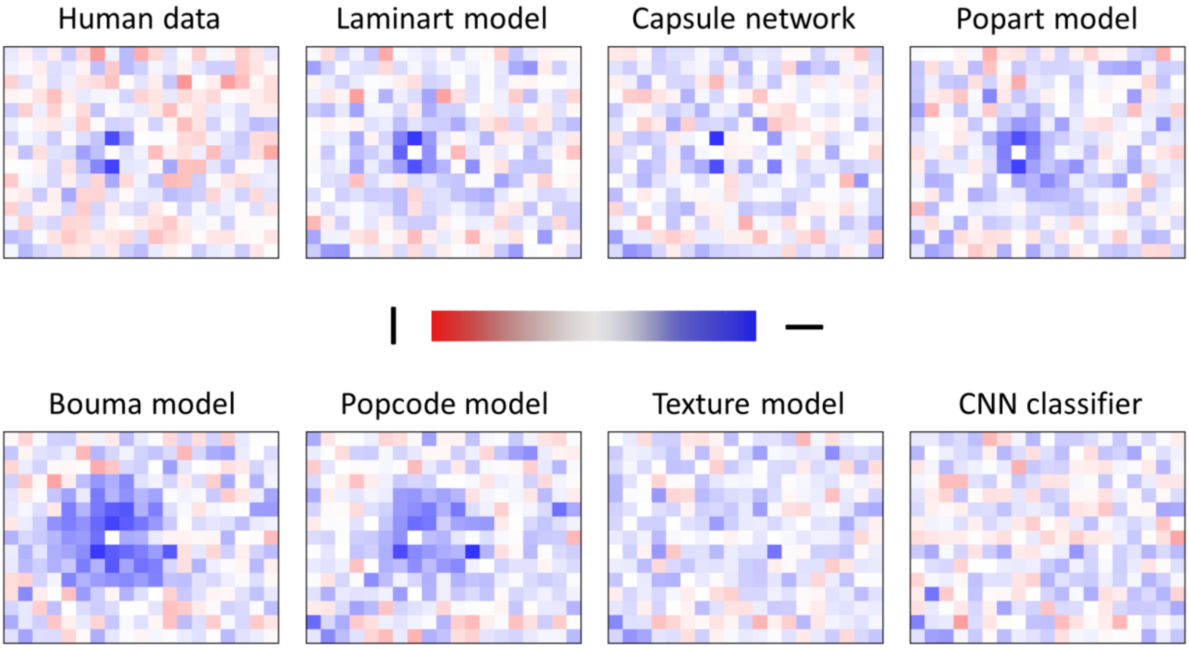

Fig A. Fine-grained version of the selection measures presented in Fig 3 (see main text). A red or a blue slot respectively indicate, for each location of the dense display, the fraction of vertical or horizontal flankers that were selected by the GA procedure, compared to randomly selected displays, after 6 generations.