

Crowding and the importance of grouping and segmentation processes in human vision

Présentée le 10 juin 2021

Faculté des sciences de la vie
Laboratoire de psychophysique
Programme doctoral en neurosciences

pour l'obtention du grade de Docteur ès Sciences

par

Alban BORNET

Acceptée sur proposition du jury

Prof. C. Petersen, président du jury
Prof. M. Herzog, directeur de thèse
Dr R. Chakravarthi, rapporteur
Prof. R. Van den Berg, rapporteur
Prof. P. Ramdya, rapporteur

Remerciements

Ces quatre dernières années se sont déroulées dans un bonheur presque absolu qu'il serait présomptueux d'attribuer uniquement à mon optimisme débordant. Merci à Naïg qui m'a rappelé que le zèle est aliénant, qu'il faut chaque jour célébrer notre existence nécessairement matérielle par la frénésie, la danse et l'ivresse. Je ne saurais imaginer quelqu'un de plus clairvoyant et sans son aura, je ne serais qu'un être fade et austère. Merci à Aurélien qui, dans un même esprit de contestation du travail huguenot, m'a soutenu que toutes les bonnes idées de sa thèse, il les a trouvées sur les chiottes. Encore aujourd'hui c'est avec lui que j'ai passé la plus grande partie de ma vie. Son calme et sa sagesse m'ont marqué à jamais.

Merci à mes parents sans qui tout ce monde n'aurait pas existé pour moi. Grâce à eux, j'ai eu l'opportunité de suivre mes choix dans un minimum de contraintes. Merci à mon Papa qui s'intéresse beaucoup au monde de la physique, en particulier la cosmologie, qui a même un jour formulé sa propre équation de l'Univers, sur le dos d'une enveloppe qui trône encore sur mon bureau. Merci à ma Maman qui apporte l'allégresse partout où elle va, qui prétend ne pas s'intéresser à la science mais sans l'admettre s'enthousiasme de la logique algébrique au vu de sa répartie percutante et de son humour sagace.

Merci à Adrien. Tout porte à croire qu'il n'est que pur intellect et art accompli, mais en réalité il est également d'une stupidité étonnante, ce qui fait de lui un homme complet. Je n'ai jamais travaillé avec lui, car je me suis toujours bien trop amusé pour appeler ça du travail. Merci à Ophélie qui a toujours été là pour désamorcer les situations les plus pesantes par des pirouettes rocambolesques, et qui est encore plus présente lorsqu'il s'agit de marquer la fin de journée, à ma grande satisfaction.

Merci à Florian et Martin, compagnons de fortune avec qui j'ai découvert les joies de la pétanque et de la désinvolture. Merci à Gizay et Dario qui étaient bien souvent les seuls à vouloir m'accompagner pour une bière post-apocalyptique. Merci à Oh-Hyeon pour sa conversation fantasque et son énergie inépuisable. Merci à Cédric pour les petits-déjeuners en amoureux et nos aventures romanesques dans les terres arides des mondes oubliés.

Merci à Greg pour m'avoir supervisé, officiellement mais méticuleusement. Merci à Mauro qui m'a aidé à finaliser un chapitre de ma thèse sans perdre patience. Merci à Ben pour avoir corrigé les (très rares, cela va de soi) fautes d'anglais. Merci à tous les membres de mon laboratoire pour l'ambiance agréable à laquelle ils ont participé et tous les événements que nous avons pu partager ensemble.

Merci à Michael. Il est difficile d'imaginer un meilleur superviseur. Outre son intelligence qui dépasse l'entendement, il recèle une grande humanité. Un jour, il a dit: "Vous savez, vous ne travaillez pas pour moi. Je travaille pour vous." Il prend son rôle de supervision très à cœur et fait tout son possible pour nous diriger vers un travail scientifique de qualité.

Enfin, merci à vous, expert, président ou simple amateur de thèse de 200 pages, qui sacrifiez une partie de votre temps pour lire le fruit de mon travail et m'aider à le rendre meilleur. J'espère que vous apprécierez la lecture.

Résumé

Traditionnellement, l'étude scientifique de la vision chez l'être humain consiste à décomposer les calculs complexes effectués par le cortex visuel en une cascade d'opérations basiques implémentées par de petits circuits neuronaux. Selon cette idée, l'information visuelle circule dans un seul sens, ou *feedforward*, comme dans une machine automatique extrêmement efficace, où l'activité des millions de neurones de la rétine est transformée en concepts visuels abstraits et complexes en à peine un claquement de doigts.

Cette approche a permis de découvrir de nombreux mécanismes fondamentaux intervenant dans le cortex visuel et d'élaborer des modèles aussi performants que de vrais humains dans de nombreuses tâches visuelles très complexes. Par exemple, les réseaux neuronaux profonds, ou *deep neural networks*, sont considérés comme des représentations fidèles du cortex visuel humain, ainsi que les meilleurs algorithmes dans le domaine de l'intelligence artificielle.

Cependant, en se basant uniquement sur des modèles *feedforward* et des mesures de l'activité de circuits neuronaux, il est possible de passer à côté d'autres aspects fondamentaux de la vision humaine, tels que l'influence des connexions récurrentes dans le cerveau ou encore l'importance du contexte global d'une image lorsqu'elle est analysée par le cortex visuel.

Des paradigmes psychophysiques peuvent être utilisés pour sonder les méandres de la vision humaine, par exemple le *visual crowding* (« encombrement visuel »), dans lequel un objet-cible est plus difficile à identifier lorsqu'il est entouré par d'autres objets qui lui ressemblent. De nombreuses expériences utilisant ce paradigme ont produit des résultats qui ne peuvent pas être expliqués par les modèles traditionnels de la vision humaine. Par exemple, en présence de nombreux objets qui se ressemblent entre eux, l'objet-cible est aussi facile à reconnaître que s'il était isolé (*uncrowding*). Nous appelons ces résultats les effets globaux du *crowding*.

Dans cette thèse, je commence par analyser quels modèles de la vision humaine sont capables de produire du *uncrowding*. Je sélectionne ces modèles en fonction de différents facteurs architecturaux et fonctionnels et je compare leur performance. Je montre que le seul modèle qui se comporte comme le système visuel humain est un modèle de segmentation, i.e., un

modèle où la manière dont les éléments visuels sont groupés entre eux influence la perception individuelle de ceux-ci.

Ensuite, je montre que les effets globaux du *crowding* ne peuvent pas émerger de modèle se basant uniquement sur des statistiques locales. Il a été proposé qu'un tel modèle, le *Texture Tiling model*, peut produire ces effets simplement parce qu'il contient de nombreuses dimensions, sans utiliser de processus de segmentation. Je teste ce modèle en me basant sur un grand nombre de résultats expérimentaux utilisant le *crowding*. Je montre que ce modèle est équivalent à un modèle de basse dimension et qu'il n'explique aucun des résultats testés.

Ensuite, je me concentre sur les *deep neural networks*, qui sont les représentants les plus performants des modèles *feedforward*, tant du point de vue de l'intelligence artificielle que celui des neurosciences. Je teste deux réseaux, *AlexNet* et *ResNet-50*, qui ont été proposés comme modèles du système visuel humain. Je montre que la manière dont ces réseaux sont construits fait qu'ils ne peuvent pas produire du *uncrowding*.

Enfin, j'utilise un algorithme génétique pour générer des stimuli en fonction de la performance de différents modèles. Le but est d'éviter de définir moi-même des stimuli qui favorisent certains modèles au départ. Je compare les stimuli qui sont produits par les modèles à ceux qui sont produits par les humains. Je montre que seuls les modèles qui incluent des processus de segmentation se comportent comme des humains.

Pris ensemble, les résultats de ma thèse mettent en évidence l'importance des processus de segmentation dans le système visuel humain. Ils démontrent que ces processus sont un ajout prometteur aux modèles traditionnels, pour mieux comprendre les mécanismes fondamentaux de la vision chez l'être humain.

Mots-clefs

Vision humaine, visual crowding, modélisation, groupement visuel, segmentation, interactions locales, contexte global, réseaux feedforward, réseaux récurrents, algorithme génétique

Abstract

Human vision has evolved to make sense of a world in which elements almost never appear in isolation. Surprisingly, the recognition of an element in a visual scene is strongly limited by the presence of other nearby elements, a phenomenon known as visual crowding. Crowding impacts vision at all levels and is thus a versatile tool to understand the fundamental mechanisms of vision.

For decades, visual crowding was perfectly well explained by traditional feedforward models of vision. In these models, vision starts with the detection of low-level features. This information is combined locally along the hierarchy of the visual cortex to build more and more complex feature detectors, until neurons respond selectively and robustly to complex objects. Crowding happens when nearby elements interfere in this local feature combination process and impair target recognition.

However, recent studies have shown that crowding is not determined by local interactions but by the global configuration across the entire visual field. Depending on how elements group together, crowding can even almost disappear, a phenomenon called uncrowding. Hence, crowding is rather a complex, global and high-level phenomenon, that simple feedforward models cannot explain.

In this thesis, I first analyse which models of crowding can explain uncrowding. I compare the performance of diverse models, selected according to different architectural and functional features, such as feedforward vs. recurrent architecture, local or global information processing, including a grouping stage or not. I show that the only model that reproduces human behaviour includes a dedicated recurrent grouping processing stage.

Second, I show that global effects in crowding cannot be explained by low-level accounts. It was argued that the Texture Tiling model, based on a complex and high-dimensional pooling stage, may account for global effects in crowding, without requiring any recurrent grouping stage. To test this model, I use a large pool of recent crowding data. I show that the Texture Tiling model is equivalent to a simple pooling model and is thus as limited as these models.

Next, I focus on deep neural networks, which are well in the spirit of the feedforward framework of vision and have become state-of-the-art models both in computer vision and neuroscience. I test whether AlexNet and ResNet-50, which have been proposed as realistic models of the visual system, exhibit uncrowding. I show that these networks do not reproduce uncrowding for principled reasons.

Finally, I use a genetic algorithm to generate stimuli based on the performance of different models, i.e., in a bottom-up manner. The goal is to avoid using stimuli that favour models of grouping from the start. I compare the distribution of stimuli that are produced by the models to the ones that are produced by humans. I show that only the models that include grouping and segmentations processes behave like humans.

Taken together, the results in my thesis highlight the importance of recurrent grouping and segmentation processes in human vision when large portions of the visual field are involved. These results can be used as direct guidelines for future models of vision, in order to constraint how recurrent processing should be incorporated to improve the performance of deep neural networks and other feedforward models of vision, and help them generalize to more complex visual inputs.

Keywords

Human vision, visual crowding, modelling, grouping, segmentation, local pooling, global processing, feedforward networks, recurrent networks

List of publications

Publications and manuscripts included in the thesis main body

- Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A.M. and Herzog, M.H., 2019. Beyond Bouma's window: How to explain global aspects of crowding? *PLoS computational biology*, 15(5), p.e1006580.
I wrote code and ran simulations for several models, wrote the description of these models and provided corrections to the manuscript.
- Bornet, A., Choung, O. H., Doerig, A., Whitney, D., Herzog, M. H. and Manassi, M. (Submitted) Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Journal of Vision*.
I conducted the underlying research, wrote most of the code, wrote the manuscript together with the last author and generated all the figures.
- Doerig, A.[†], Bornet, A.[†], Choung, O.H. and Herzog, M.H., 2020. Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research*, 167, pp.39-45.
[†]Equal contributions
I conducted the underlying research, wrote the code and wrote the manuscript together with the other first author.
- Bornet, A.[†], Doerig, A.[†], Herzog, M.H. and Van der Burg, E. (submitted). Shrinking Bouma's window: Models of Crowding in Dense Displays. *PLoS computational biology*.
[†]Equal contributions
I conducted the underlying research, wrote all the code and most of the manuscript.

Other publications

- Bornet, A., Kaiser, J., Kroner, A., Falotico, E., Ambrosano, A., Cantero, K., Herzog, M.H. and Francis, G., 2019. Running large-scale simulations on the NeuroRobotics Platform to understand vision-the case of visual crowding. *Frontiers in neurorobotics*, 13, p.33.
I conducted the underlying research, wrote most of the code and of the manuscript.

- Choung O.H., Bornet A., Doerig A., Herzog M.H. (submitted). Dissecting (un)crowding. *Journal of Vision*.
I ran simulations for one model, wrote the model description and provided corrections to the manuscript.
- Bornet A., Doerig A., Rashal E., Herzog M.H. (in preparation). Uncrowding and attention. *In this study, we investigate whether there is an interaction between uncrowding and top-down attention. If not, uncrowding may originate from bottom-up cues, such as salient regions in the flanker configurations. This does not rule out the role of recurrent processing. I presented preliminary results of this study during ECVP 2018.*
- Bornet A., Doerig A., Herzog M.H. (in preparation). Models of feature integration in time, using the SQM paradigm.
In this study, we use data from the SQM paradigm to model feature integration in discrete time-windows. We propose to include axonal delays in a multi-layered predictive coding model to exhibit long periods of integration. I describe this preliminary work in the general discussion.

Conference proceedings

- Bornet, A., Kroner, A., Kaiser, J., Scholz, F., Francis, G., Herzog, M.H. Using the Neurorobotics Platform to explain global processing in visual crowding. *European Conference on Visual Perception (ECVP), 2018*
I conducted the underlying research and presented the talk.
- Bornet, A., Doerig, A., Van der Burg, E., Herzog, M.H. Shrinking Bouma's window: visual crowding in dense displays. *European Conference on Visual Perception (ECVP), 2019*
I conducted the underlying research and presented the talk.

I was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 720270 (Human Brain Project SGA1), No. 785907 (Human Brain Project SGA2) and No. 945539 (Human Brain Project SGA3). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscripts.

Table of content

Remerciements	2
Résumé	4
Abstract	6
List of publications	8
Table of content	10
Introduction.....	13
The traditional framework of human vision	14
Importance of specific and well-controlled probes.....	17
Visual crowding and traditional models of vision	19
Challenges to vision models and importance of visual grouping.....	22
Overview of the thesis.....	25
References	27
Chapter 1: Beyond Bouma’s window - How to explain global aspects of crowding?	33
Abstract.....	34
Author Summary	35
Introduction	36
Methods.....	41
Results.....	43
Discussion	45
Model comparison	45
Future Models.....	47
Conclusion.....	49
References	50
Chapter 2: Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing	53
Abstract.....	54
Introduction	55
General Materials and Methods.....	57
Mongrel generation	57
Ethics.....	57
TTM & Grouping Effects.....	58
Methods	58
Results	60
TTM & prediction power	65
Grouping effects & target cueing.....	66

TTM & Face Crowding	69
Methods	69
Results	72
Discussion	78
TTM & grouping effects	78
TTM & face crowding.....	80
Model assessment method	82
Model improvements	83
References	85
Chapter 3: Crowding reveals fundamental differences in local vs. global processing in humans and machines...	89
Abstract.....	90
Introduction	91
Methods	95
Experiment 1a	95
Experiment 1b.....	96
Experiment 2	96
Results	98
Experiment 1a	98
Experiment 1b.....	98
Experiment 2	100
Discussion	102
References	105
Chapter 4: Shrinking Bouma’s window - Models of crowding in dense displays	108
Abstract.....	109
Author summary.....	110
Introduction	111
Methods	116
Results	119
Pooling models.....	119
Grouping models.....	120
Discussion	123
References	126
General discussion	129
Summary of the results	130
Limitations	137
Prospects.....	138
Conclusion.....	143
References	145

Supplementary information	150
A: Supplementary information for Chapter 1	151
SA1: Epitome model	151
SA2: Single texture model	152
SA3: Texture tiling model (TTM)	153
SA4: Deep textures	155
SA5: Wilson & Cowan network with end-stopped receptive fields	157
SA6: Zhaoping's V1 recurrent model	159
SA7: A variation of the Laminart model	160
SA8: Alexnet (convolutional neural network).....	164
SA9: Hierarchical sparse selection (HSS).....	166
SA10: Saccade-confounded summary statistics	168
SA11: Population coding.....	168
SA12: Fourier model	169
B: Supplementary information for Chapter 2	171
SB1: Comparison between Lines and Completion experiments	171
SB2: Shapes experiment with diamonds	172
SB3: Shapes and Patterns experiments with larger fovea parameter	173
SB4: Butterflies experiment.....	174
SB5: TTM and prediction power - Template match algorithm performance	174
SB6: TTM and prediction power - Separate experiments.....	175
SB7: Pointers location in Manassi et. al (2012)	175
SB8: Effect of pointers in the TTM	176
SB9: Single face discrimination task - reverted back	177
SB10: Mongrel gender matching algorithm	177
SB11: TTM & Pixel density - Detailed methods	178
C: Supplementary information for Chapter 4	179
SC1: Bouma's law model	179
SC2: Population coding model	180
SC3: Texture model.....	182
SC4: CNN classifier	184
SC5: Contour segmentation model ("Laminart")	185
SC6: Capsule network	187
SC7: Two-stage model ("Popart")	189
SC8: Human experiment for proportion measure	190
References	192
Curriculum vitae	195

Introduction

The traditional framework of human vision

Humans recognize objects without effort, even though there are thousands object classes (e.g. car, plant, dog, bottle, etc.), and there are virtually infinitely many possible instances for each object (different shapes, sizes, colors, poses, internal configurations, locations, points of view, lighting conditions, etc.). In order to solve this task, the human visual system must convey information from the activity of more than 10 million cone photoreceptors per retina (1) into robust and invariant object-like representations. Although it seems extremely complex in terms of computations, this task requires no more than 150 ms of cortical processing in humans (2–4). How is this even possible? What is the machinery underlying such an efficient system?

To unveil the mechanisms of human vision, it is necessary to rely on simple (yet powerful) models. To this end, based on electro-physiological and anatomical studies of different regions of the visual cortex in cats and monkeys (5–7), vision was first formulated as a feedforward and hierarchical process. Object recognition was shown to occur in a dedicated stream of the visual cortex (ventral stream; 8), organized in subsequent cortical areas whose role is to encode different features that characterize the content of the visual input (9,10; see Fig 1, top). Connections between the layers only go in one direction: from simple, very localized and retinotopically organized features in the first visuo-cortical areas (11) to more complex and abstract features in the higher regions of the ventral stream, that respond specifically to objects but are relatively oblivious to low-level properties of the visual input (12). Importantly, the neurons' receptive field size increases along the processing hierarchy, and their responses become more robust and invariant to continuous transformation of objects in the visual field (pose, size, shape, etc.).

Based on the incoming activity of the retinal ganglion cells, the first layers of the visual cortex detect simple, local and low-level oriented edges (Fig 1, bottom). Then, the subsequent layers combine this information locally and in a feedforward manner to build more and more complex feature detectors (e.g., object parts, such as doorknobs or wheels). This feature combination process leads to a final stage in which the patterns of activity are selective and robust to the presence of complex objects (e.g., any type of car, from any point of view), which allows complex object recognition.

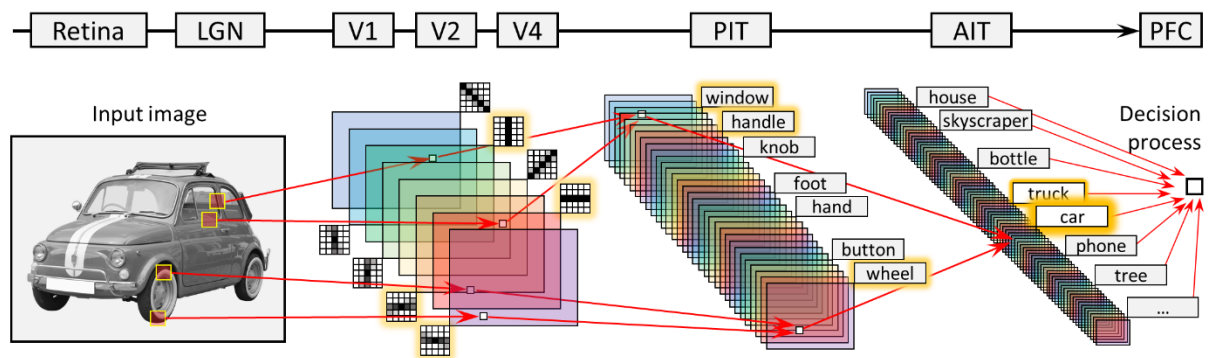


Fig 1. Top. Extremely simplified schematic of the ventral stream anatomy, as observed in the primate visual cortex. Light coming from the outer world is first processed by the retina and sent to the visual cortex through the visual part of the thalamus, the Lateral Geniculate Nucleus (LGN). The early layers of the visual cortex (V1, V2, V4) encode the content of the image less and less locally, as receptive field sizes increase with cortical depth. Neurons in the infero-temporal cortex (IT) respond with high selectivity to different object parts (posterior IT; PIT) and whole objects (anterior IT; AIT) but are oblivious to simple and local image features. Finally, object-related information coming from IT is integrated in the Pre-Frontal Cortex (PFC) to make a decision regarding the high-level content of the visual input. **Bottom.** How object recognition is supposed to work according to traditional feedforward models of human vision. The processing architecture loosely matches the anatomy of the ventral stream depicted above. Between layers, information is filtered and integrated into more and more complex features, using simple and local mathematical transformations, such as local filter convolutions, until complex objects are detected, selectively and robustly.

Modelling visual perception as a feedforward process is particularly useful since complex and high-level visual processing, such as object recognition, can be broken down into local and mathematically tractable sub-problems, performed by simple neuronal circuits (13). In past decades, electro-physiological studies discovered various low-level basic input-output functions in the early visual cortex and described them in terms of local synapse circuitries, e.g., surround suppression (11,14,15), cross-orientation suppression (16,17), or end-stopping (18–20). All these functions are easy to implement and to understand, because they can be reproduced by simple neuronal circuits.

According to traditional vision research, it is sufficient to combine these basic processing blocks in a hierarchical, retinotopic and feedforward fashion to explain how humans recognize objects in a scene (21,22). As a major breakthrough, it was shown that object representations invariant to any transformation can be built with local feature detectors and max-pooling operations (21,23). Based on this idea, many artificial models of object recognition were designed and

tested (24–29). All these models reach high levels of performance in object recognition tasks, while relying on a small number of parameters and computations.

Following the same inspiration, more recently and thanks to the increasing capacity of modern computers, a new class of models, namely the deep feedforward convolutional neural networks, were developed (30,31). These networks can not only perform object recognition as well as humans (32–35), but they also excel in a large number of vision-based tasks, often exceeding human performance. These tasks range from object segmentation (36), to image synthesis (37,38) and scene understanding (39).

Importantly, the traditional framework of human vision is particularly well embedded in deep convolutional neural networks: they use the same feedforward layered architecture and the same kind of basic operations to build linearly separable object-like representations. In addition, deep neural networks share interesting similarities with the human visual system. For example, after being trained on image recognition, activity in the different layers of deep neural networks resembles the activity observed in the visual areas of the primate ventral stream (40–43). The same applies to the shape of their receptive fields (44,45). For all these reasons, deep neural networks have also been studied as models of human vision (46,47).

Importance of specific and well-controlled probes

Reducing vision to basic computations performed by local cortical circuits makes electrophysiological measurements the natural choice to access fundamental properties of the visual cortex. Most variables and sources of noise are controlled, allowing to access local circuits connectivity with extreme precision. However, to study these circuits in isolation, feedback and lateral connections must play only a little role in the visual cortex, such as modulatory effects (48–50), noise disambiguation (51–53), uncertainty reduction (54), or the propagation of learning signals as in biologically plausible deep neural networks (55–58).

It has been shown that this is not the case. For example, even at the very first stage of the visual cortex (V1), 85% of the input comes from intracortical regions, i.e., from lateral or top-down connections (59). Moreover, characterizing orientation tuning properties of macaques V1 neurons in terms of linear response to simple stimuli has little predictive power for their response to natural images (60,61). The same holds for hue tuning curves measured in the V4 area. Hue tuning properties in response to artificial stimuli are considerably altered in natural images (62). This suggests that it is almost impossible to study even the most basic circuits of the visual cortex in isolation without being restricted to extremely simple paradigms, measurements and models. Finally, the complexity associated to summarizing all physiological measurements of any functional area of the human visual cortex under a comprehensive theory increases exponentially with the number of measured neurons (63).

These findings are in sharp contrast with findings in the retina, where neural firing to natural scenes is captured well by convolutional neural network models, even at the cellular level (64); or with the LGN, whose cell responses to complex image patterns are well explained by a feedforward linear filter model (65). The difference with the visual cortex is that recurrent connections play a sparser role in the retina and the LGN. Hence, the expected output of the retina and the LGN are well-defined and formulating them as input/output functions is easier. In the visual cortex, it is harder to define what or where the output is: visual information is not transient, but is integrated for long periods of time (66–69), up to 450 ms. The only relevant outputs that can be defined without ambiguity are high-level, human-level perceptions and sensations. Hence, it is very hard, if not impossible, to map different sets of physiological

measurements to actual functions of the visual cortex, given the complex ontology and numerosity of its possible outputs (70).

According to Marr's tri-level framework (71–73), the visual cortex can be studied at three different levels: computational (what are the goals pursued by human vision and why are they appropriate), algorithmic (how are computations implemented and what do they represent), and hardware implementation (how can these algorithms be realized physically). From that perspective, physiological measurements are powerful tools to addressing mainly implementational questions about the visual cortex but seem less adapted to algorithmic and computational questions. For all these reasons, it is crucial to pair physiological observations with measurements that are taken at a higher level of abstraction, namely, psychophysical paradigms (22). Psychophysics bridges the gap between the stimulation of the visual system and high-level processes by providing well-controlled procedures to access actual human perceptions and sensations (74–76). Pairing physiological and psychophysical measurements is not new. For example, evidence for the existence of invariant object-representations in the visual cortex were backed-up by numerous psychophysical studies (77–79).

Compared to physiological measurements, psychophysical paradigms have the possibility to ask different questions that may sometimes be more adapted to the required level of understanding. Rather than being in competition, these different levels complement and constrain each other. Eventually, to fully understand human vision, we need a coherent set of theories at all levels. Importantly, psychophysical measurements are less stringently bound by stimulus and paradigm complexity because they reside on a higher level of abstraction. Moreover, compared to physiological measurements, it is just as straightforward to use psychophysical paradigms to validate models of the visual system, since they rely on a set of well-controlled assessment methods and well-defined stimuli. This would not be the case, for example, for psychological paradigms. As a probe into the fundamental processes of human vision, the psychophysical paradigm that was used through this thesis is visual crowding.

Visual crowding and traditional models of vision

In visual crowding, identification of a target is impaired by the presence of nearby flankers. The target is visible, but its features appear jumbled and distorted (Fig 2a). Crowding is ubiquitous. It occurs for simple lines (80), letters (81), digits (82), Gabors (83), faces (84), everyday objects (85), and is observed in other modalities, e.g., audition (86–89) or touch (90). Because it affects so many aspects of perception, crowding may be the burden that any perceptual system needs to cope with because of its constraints (finite size, time, resolution, etc.). It is thus a precious tool to study the fundamental mechanisms of human visual perception (91). In the last decades, visual crowding has been characterized in detail by numerous studies that have focused mainly on very simple paradigms, i.e., involving few flankers only.

Based on the result of these studies, several hallmarks of crowding were formulated. Crowding was found to be stronger for flankers sharing similar low-level features with the target, such as orientation, colour, size, etc. (80,92–94). The spatial extent of crowding was determined to grow linearly with eccentricity, i.e., equal to half the target eccentricity when flankers are aligned in the radial direction (Bouma's law; see Fig 2b; 78,92) and 2-3 times smaller when aligned in the tangential direction (radial-tangential anisotropy; 93). Crowding was shown to be stronger with flankers on the peripheral side of the target than on the foveal side (inward-outward anisotropy; 92,94,95), and weaker in the lower than in the upper visual field (99).

The majority of models of visual crowding proposed that crowding is the consequence of feature integration or pooling (100–107). This explanation is in line with feedforward models of vision, in which features must be integrated along the visual processing hierarchy to yield target identification. Crowding happens when the target's features are compulsorily pooled with features of the flankers because they fall within the same receptive field (see Fig 2c-d). The perceived features of the target are averaged with the flankers' perceived features. Importantly, pooling is thought to happen at the early stages of vision. This means that even though crowding happens between faces or houses, it is the result of the interactions between low-level features that make up the objects (108). Depending on the paradigm, crowding may be partially imputed to substitution errors, in which the representation of the spatial order of elements is noisy (109–111). In this case, the reported visual element is not the target but one of the flankers, leading to more errors.

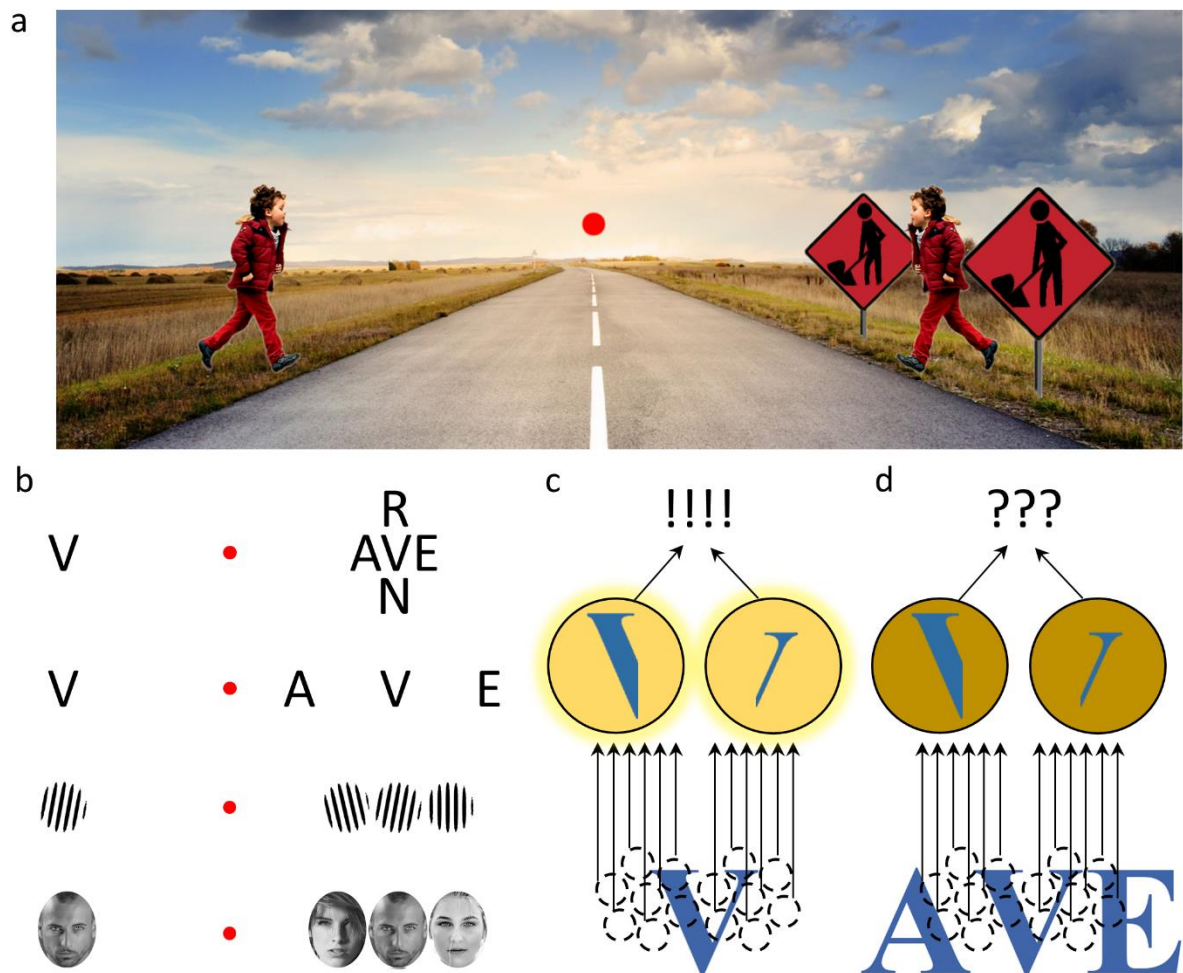


Fig 2. a. Visual crowding in everyday life. The child on the right is harder to identify compared to the one on the left because of the neighbouring signs that share similar features (colours, shape, size). **b.** Crowding in psychophysical experiments. Crowding happens in various visual recognition tasks (letters, Gabors, faces, etc.). The target letter is easier to identify on the right than on the left, except at the second row, where the flankers lie outside Bouma's window. **c.** Target identification according to feedforward models of vision. The output of local contrast detectors, the receptive fields of which are represented by the small black circles, are integrated in an intermediate pooling stage (black arrows). The output of this integration is used at a higher-level to recognize the letter (exclamation marks). **d.** Visual crowding according to feedforward models of vision. When flankers are added, some irrelevant information is pooled along the processing hierarchy. As a consequence, the neurons that integrate local features along the hierarchy are less optimally activated than in the unflanked condition. Hence, it is harder for the higher-level neurons to identify the target (question marks).

fMRI studies located the cortical area where visual crowding happens at least after the V1 level (112), which matches the pooling theory, since feature detection (the step prior to feature integration) happens in V1 (6). Moreover, the pooling explanation accounts for Bouma's law, since receptive field sizes grow approximately linearly with eccentricity (cortical magnification;

110–113). Several neuroanatomical and neurophysiological studies designated area V4 as a promising locus for visual crowding. This is in quantitative agreement with Bouma's law (117), since the cortical magnification factor in this area is roughly equal to 0.5 (117,118). The shape of the receptive fields in area V4 is consistent with the radial-tangential anisotropy (118). Moreover, V4 is a plausible site for the type of feature integration that is presumed to yield crowding (119,120).

These considerations resulted in models that can be tested quantitatively. For example, Van den Berg et al. (107) proposed a model based on the spatial integration of population coding signals. This model is physiologically plausible and accounts quantitatively for different hallmarks of crowding, such as Bouma's law, the radial-tangential anisotropy or the inward-outward anisotropy. Moreover, this model explains both types of errors in crowding, substitution and averaging, using a single pooling mechanism.

To sum up, the majority of visual crowding studies are perfectly in line with the traditional framework of vision research and are well explained by local feedforward models based on pooling.

Challenges to vision models and importance of visual grouping

Recent studies have measured visual crowding using more complex flanking patterns. Importantly, the results of these studies undermine the success of feedforward models of crowding. For example, in Manassi et al. (121), human participants were asked to discriminate between a left or a right vernier target presented in the periphery, and surrounded by different flanker configurations (see Fig 3a). The task is easy when the target is alone (red dashed line) but hard when a square flanker is added (crowding; 1st column). However, with added squares, performance recovers almost to the unflanked level, an effect called *uncrowding* (2nd to 4th columns). This is in contradiction with the pooling explanation of crowding, in which adding flankers always increases crowding strength.

Uncrowding had already been measured in previous studies (122–124). However, testing more complex flanking patterns revealed that crowding strength is determined by the global configuration of flankers, i.e, in the whole visual field (5th to last columns), far beyond the range predicted by Bouma's law (Bouma's window; 118,122). In these studies, the low-level perception of a few arcmins vernier offsets is determined by high-level configuration changes occurring in a radius of almost 10 degrees (see Fig 3b). Moreover, it was shown that feature similarity between the target and the flankers is not always sufficient nor necessary to produce crowding (126).

Furthermore, crowding studies in which the flankers not only cover large portions of the visual field, but are also arranged in a dense fashion (dense displays; 124–126), produced results that are not in line with classic models of crowding. In dense displays, the range at which flankers have an influence on target discrimination performance was found to shrink to the nearest neighbour distance, way beneath Bouma's window (130). Crucially, the range of interaction does not scale with target eccentricity in dense displays (131). These results are in contradiction with Bouma's law and the pooling account of crowding.

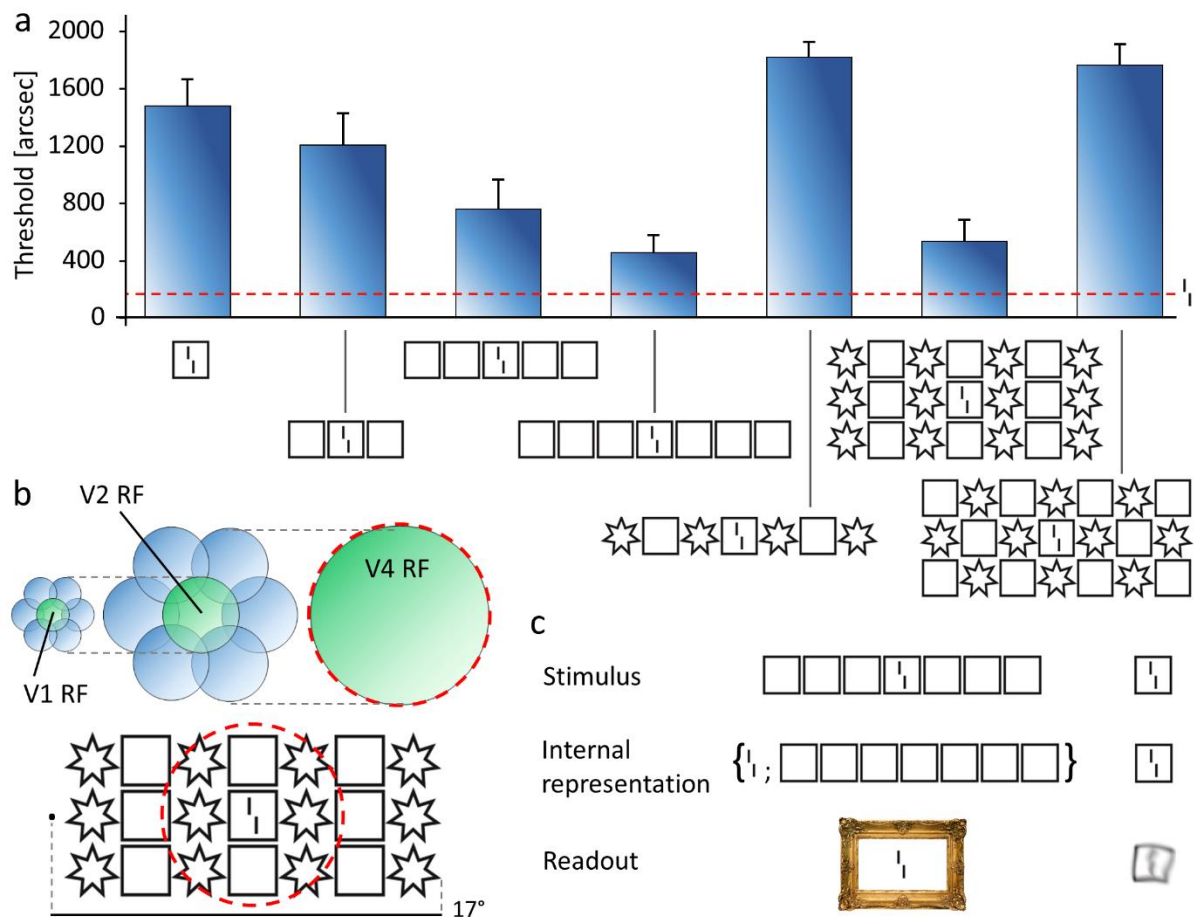


Fig 3. a. Observers were asked to discriminate between a left or right vernier presented at 9° of eccentricity and surrounded by different flanker configurations (121,125). The y-axis shows the vernier offset threshold for 75% of correct responses. When the target is alone, performance is good (red dashed line). When a square is placed around the target, performance decreases dramatically (crowding, 1st column). When more squares are added, performance recovers almost to the unflanked level (uncrowding, 2nd to 4th column). Crowding strength is strongly affected by the configuration of flankers in the whole visual field (5th to last columns). **b.** Illustration of different receptive field sizes at 9° of eccentricity. The dashed circles indicate Bouma's window. Elements having a dramatic influence on crowding strength lie far beyond Bouma's window. **c.** Two-stage model. The stimulus is first parsed into different groups, and interaction happens only within each group. When more squares are added, it becomes easier to segment the target from the group of flankers.

Moreover, studies involving displays with realistic and semantically rich content showed that crowding is not merely due to the integration of the low-level features that make up the objects but can happen at any stage of the visual hierarchy, including high-level (132,133). For example, crowding between visual scenes was dramatically reduced by removing semantic information of the flanker scenes (134). Alternatively, crowding between faces was shown to be stronger for upright than for inverted flanker faces, only for upright target faces (135,136).

Crowding and uncrowding seem to be strongly affected by the global configuration of the elements in the visual field, even at the scale of a vernier target. Local feedforward models cannot account for these effects. Previous studies have shown that crowding happens at a level where configuration information is already extracted (137) and that crowding is stronger within, rather than between, Gestalt clusters of visual elements (138). Van den Berg et al. (107) proposed that adding grouping processes to their population coding model using feedback connections could account for configuration effects in crowding. They suggested that adding inhibitory connections from higher levels in their model might suppress the integration of signals between different perceptual groups.

Moreover, the classic hallmarks of crowding can also be explained as the consequence of processes that do not have to be feedforward. For example, it was proposed that the limited resolution of selective attention is responsible for visual crowding (139). This process has a coarser grain than pure visual acuity and could, in theory, account for Bouma's law. Moreover, attention models were linked to the time resolution of crowding (140) and to stronger effects in the upper visual field (99).

For all these reasons, it is appealing to explain high-level configurational effects of crowding with models that process visual inputs globally and incorporate recurrent connections. One successful idea has been to model crowding as a two-stage process in which elements are first parsed into different groups, before interactions happen within each group only (see Fig 3c). Along these lines, Francis et al. (141) proposed a model in which elements are grouped by illusory contours and parsed in different layers of the network by a recurrent segmentation stage, reproducing the uncrowding effect. The segmentation process arises from a competition between different populations, in which activity is modulated by high-level inhibitory signals, as suggested by Van den Berg et al. (107). The success of the model of Francis et al. (141) unveils the potential importance of grouping and segmentation processes in human vision, once large portions of the visual field are taken into account.

Overview of the thesis

In this thesis, extensive modelling studies are performed to outline the importance of grouping and segmentation processes in visual crowding and human vision in general. The main approach is to pit different classes of models against each other by comparing their abilities to explain the global aspects of crowding described in the previous section. The first class of models are feedforward models, which are built upon the traditional framework of vision research. In addition, recurrent models that include grouping and segmentation processes are considered. This thesis consists of four chapters, each addressing an important aspect of crowding and human vision from a modelling perspective.

First, although Francis et al. (141) explained global aspects of crowding with a recurrent model of grouping and segmentation, it is important to make sure that feedforward models cannot. Indeed, many different feedforward models of crowding exist, all with their specific hypotheses about human vision. Moreover, other recurrent models exist that process global aspects of the visual input. For this reason, in Chapter 1, we compare the performance of different models on a large battery of crowding stimuli for which we know that global configuration plays a role in humans. We relate the success or failure of the models to their key characteristics, such as having a feedforward vs. recurrent architecture, processing information locally vs. globally, or whether they include a grouping stage or not.

Second, Rosenholtz et al. (142) suggested that global aspects of crowding could be explained by more sophisticated pooling models, such as the Texture Tiling Model (TTM). Contrary to simple pooling models, the pooling stage of the TTM is high-dimensional and preserves rich information, which supports a fine-grained representation of the visual input. This fine-grained representation could drive the global effects observed in crowding at a later post-perceptual stage, without requiring any grouping stage in the visual hierarchy. For example, Rosenholtz et al. (142) argued that uncrowding was simply caused by the reduction in target location uncertainty in the presence of many flankers (cueing). However, the model's predictions in Rosenholtz et al (142) were not tested quantitatively and very few conditions were presented. For example, many conditions in which cueing increases and crowding increases as well were omitted. For this reason, in Chapter 2, we test this model extensively using crowding paradigms

from different studies in which effects of configuration are observed in humans (121,125,126,143).

Third, deep feedforward convolutional neural networks have become state-of-the-art models both in computer vision and neuroscience. It has been shown that these networks are subject to crowding (144). In Chapter 3, we investigate whether they also reproduce uncrowding, as observed in humans. We test different versions of deep networks that have been proposed as models of the visual system, namely AlexNet (31) and ResNet-50 (35). Moreover, we test a version of ResNet-50 that has been trained to focus on global aspects of the visual input by removing textural information in the training set (145).

Fourth, it is important to test the predictions of grouping and segmentation processes beyond uncrowding paradigms. In Chapters 1 to 3, stimuli are “cherry-picked” to pit models against each other. However, a fairer model comparison would involve stimuli that are not designed to highlight the importance of grouping processes. To this end, in Chapter 4, we used the data of Van der Burg et al. (130) in which the range of interaction between visual elements is shrunk to the nearest neighbour distance, which is the exact opposite as in uncrowding paradigms (121,125). Importantly, the stimuli in this paradigm are generated using a genetic algorithm (146), i.e. in a bottom-up manner, which means that they are not chosen by the modeller.

To give an overview of the main contributions of this thesis, I first show that the only models of crowding that are able to explain uncrowding include a dedicated recurrent grouping stage (Chapter 1). Second, I show that global effects in crowding cannot be explained by low-level accounts, even when including a high-dimensional pooling stage (Chapter 2). Third, I show that deep feedforward convolutional networks do not reproduce uncrowding for principled reasons (Chapter 3). Finally, I show that grouping and segmentation processes are crucial to explain human behaviour beyond uncrowding paradigms (Chapter 4). Taken together, the results in my thesis highlight the importance of recurrent grouping and segmentation processes in human vision when large portions of the visual field are involved. These results can be used as direct guidelines for future models of vision, in order to constraint how recurrent processing should be incorporated in deep neural networks to improve their performance and help them generalize to more complex visual inputs.

References

1. Curcio CA, Sloan KR, Kalina RE, Hendrickson AE. Human photoreceptor topography. *J Comp Neurol*. 1990;292(4):497–523.
2. VanRullen R. The power of the feed-forward sweep. *Adv Cogn Psychol*. 2007;3(1–2):167.
3. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *nature*. 1996;381(6582):520–2.
4. VanRullen R, Thorpe SJ. Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*. 2001;30(6):655–68.
5. 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.
6. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*. 1962;160(1):106.
7. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *J Physiol*. 1968;195(1):215–43.
8. Ungerleider LG, Haxby JV. 'What' and 'where' in the human brain. *Curr Opin Neurobiol*. 1994;4(2):157–65.
9. Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. In: *Cerebr cortex*. Citeseer; 1991.
10. Van Essen DC. Visual areas of the mammalian cerebral cortex. *Annu Rev Neurosci*. 1979;2(1):227–61.
11. Sceniak MP, Ringach DL, Hawken MJ, Shapley R. Contrast's effect on spatial summation by macaque V1 neurons. *Nat Neurosci*. 1999;2(8):733–9.
12. Bruce C, Desimone R, Gross CG. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J Neurophysiol*. 1981;46(2):369–84.
13. Barlow HB. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*. 1972;1(4):371–94.
14. Knierim JJ, Van Essen DC. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J Neurophysiol*. 1992;67(4):961–80.
15. Stuart JA, Burian HM. A study of separation difficulty*: Its relationship to visual acuity in normal and amblyopic eyes. *Am J Ophthalmol*. 1962;53(3):471–7.
16. Carandini M, Heeger DJ, Movshon JA. Linearity and normalization in simple cells of the macaque primary visual cortex. *J Neurosci*. 1997;17(21):8621–44.
17. Morrone MC, Burr DC, Maffei L. Functional implications of cross-orientation inhibition of cortical visual cells. I. Neurophysiological evidence. *Proc R Soc Lond B Biol Sci*. 1982;216(1204):335–54.
18. Dobbins A, Zucker SW, Cynader MS. Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature*. 1987;329(6138):438–41.
19. DeAngelis GC, Freeman RD, Ohzawa I. Length and width tuning of neurons in the cat's primary visual cortex. *J Neurophysiol*. 1994;71(1):347–74.
20. Gilbert CD. Laminar differences in receptive field properties of cells in cat primary visual cortex. *J Physiol*. 1977;268(2):391–421.
21. Perrett DI, Oram MW. Neurophysiology of shape processing. *Image Vis Comput*. 1993;11(6):317–33.
22. DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? *Neuron*. 2012;73(3):415–34.
23. Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: *Competition and cooperation in neural nets*. Springer; 1982. p. 267–85.
24. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci*. 1999;2(11):1019–25.
25. 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.
26. Wallis G, Rolls ET. Invariant face and object recognition in the visual system. *Prog Neurobiol*. 1997;51(2):167–94.

27. Ullman S, Basri R. Recognition by linear combination of models. MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB; 1989.
28. Hummel JE, Stankiewicz BJ. An architecture for rapid, hierarchical structural description. *Atten Perform XVI Inf Integr Percept Commun*. 1996;93–121.
29. Van Rullen R, Gautrais J, Delorme A, Thorpe S. Face processing using one spike per neurone. *Biosystems*. 1998;48(1–3):229–39.
30. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278–324.
31. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
32. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ArXiv Prepr ArXiv14091556*. 2014;
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. p. 1–9.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770–8.
36. Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K. Detectron. 2018.
37. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: *Advances in neural information processing systems*. 2014. p. 2672–80.
38. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. *ArXiv Prepr ArXiv181204948*. 2018;
39. 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.
40. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol*. 2014;10(11):e1003915.
41. Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, et al. Task-Driven Convolutional Recurrent Models of the Visual System. *ArXiv Prepr ArXiv180700053*. 2018;
42. 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.
43. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 586–95.
44. 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;
45. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Springer; 2014. p. 818–33.
46. Kietzmann TC, McClure P, Kriegeskorte N. Deep neural networks in computational neuroscience. *BioRxiv*. 2018;133504.
47. VanRullen R. Perception science in the age of deep neural networks. *Front Psychol*. 2017;8:142.
48. Büchel C, Friston KJ. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cereb Cortex N Y NY* 1991. 1997;7(8):768–78.
49. Lamme VA, Super H, Spekreijse H. Feedforward, horizontal, and feedback processing in the visual cortex. *Curr Opin Neurobiol*. 1998;8(4):529–35.
50. Hupé JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*. 1998;394(6695):784–7.
51. Mumford D. On the computational architecture of the neocortex. *Biol Cybern*. 1992;66(3):241–51.
52. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. *JOSA A*. 2003;20(7):1434–48.

53. George D, Hawkins J. A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In: Proceedings 2005 IEEE International Joint Conference on Neural Networks, 2005. IEEE; 2005. p. 1812–7.
54. Chikkerur S, Serre T, Tan C, Poggio T. What and where: A Bayesian inference theory of attention. *Vision Res.* 2010;50(22):2233–47.
55. Sacramento J, Costa RP, Bengio Y, Senn W. Dendritic error backpropagation in deep cortical microcircuits. *ArXiv Prepr ArXiv180100062.* 2017;
56. Lillicrap TP, Cownden D, Tweed DB, Akerman CJ. Random synaptic feedback weights support error backpropagation for deep learning. *Nat Commun.* 2016;7(1):1–10.
57. Whittington JC, Bogacz R. Theories of error back-propagation in the brain. *Trends Cogn Sci.* 2019;23(3):235–50.
58. Liao Q, Leibo JZ, Poggio T. How important is weight symmetry in backpropagation? *ArXiv Prepr ArXiv151005067.* 2015;
59. Olshausen BA, Field DJ. What is the other 85 percent of V1 doing. Van Hemmen T Sejnowski Eds. 2006;23:182–211.
60. David SV, Gallant JL. Predicting neuronal responses during natural vision. *Netw Comput Neural Syst.* 2005;16(2–3):239–60.
61. David SV, Vinje WE, Gallant JL. Natural stimulus statistics alter the receptive field structure of v1 neurons. *J Neurosci.* 2004;24(31):6991–7006.
62. Benjamin AS, Ramkumar P, Fernandes H, Smith MA, Kording KP. Hue tuning curves in V4 change with visual context. *bioRxiv.* 2019;780478.
63. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. *Nat Neurosci.* 2019;22(11):1761–70.
64. McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S. Deep learning models of the retinal response to natural scenes. In: *Advances in neural information processing systems.* 2016. p. 1369–77.
65. Dan Y, Atick JJ, Reid RC. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci.* 1996;16(10):3351–62.
66. Drissi-Daoudi L, Doerig A, Herzog MH. Feature integration within discrete time windows. *Nat Commun.* 2019;10(1):1–8.
67. Otto TU, Ögmen H, Herzog MH. The flight path of the phoenix—The visible trace of invisible elements in human vision. *J Vis.* 2006 Aug 1;6(10):7–7.
68. Otto TU, Ögmen H, Herzog MH. Feature integration across space, time, and orientation. *J Exp Psychol Hum Percept Perform.* 2009;35(6):1670–86.
69. Lee TS, Nguyen M. Dynamics of subjective contour formation in the early visual cortex. *Proc Natl Acad Sci.* 2001;98(4):1907–11.
70. 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.
71. Marr D. *Vision: A computational investigation into the human representation and processing of visual information.* 1982;
72. Bickle J. Marr and reductionism. *Top Cogn Sci.* 2015;7(2):299–311.
73. McClamrock R. Marr’s three levels: A re-evaluation. *Minds Mach.* 1991;1(2):185–96.
74. Gescheider GA. *Psychophysics: the fundamentals.* Psychology Press; 2013.
75. Strasburger H. Software for visual psychophysics: an overview. *VisionScience Com.* 1995;
76. Treutwein B. Adaptive psychophysical procedures. *Vision Res.* 1995;35(17):2503–22.
77. Bülthoff HH, Edelman S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc Natl Acad Sci.* 1992;89(1):60–4.
78. Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. *Curr Biol.* 1995;5(5):552–63.
79. Tarr MJ. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychon Bull Rev.* 1995;2(1):55–82.

80. Andriessen JJ, Bouma H. Eccentric vision: Adverse interactions between line segments. *Vision Res.* 1976 Jan 1;16(1):71–8.
81. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Res.* 1973;13(4):767–82.
82. 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.
83. Solomon JA, Felisberti FM, Morgan MJ. Crowding and the tilt illusion: Toward a unified account. *J Vis.* 2004;4(6):9–9.
84. Martelli M, Majaj NJ, Pelli DG. Are faces processed like words? A diagnostic test for recognition by parts. *J Vis.* 2005;5(1):6–6.
85. Wallace JM, Tjan BS. Object crowding. *J Vis.* 2011;11(6):19–19.
86. Plack CJ, Carlyon RP, Viemeister NF. Intensity discrimination under forward and backward masking: role of referential coding. *J Acoust Soc Am.* 1995;97(2):1141–9.
87. Zeng F-G, Turner CW. Intensity discrimination in forward masking. *J Acoust Soc Am.* 1992;92(2):782–7.
88. Oberfeld D. The mid-difference hump in forward-masked intensity discrimination. *J Acoust Soc Am.* 2008;123(3):1571–81.
89. Jones MR, Boltz M, Kidd G. Controlled attending as a function of melodic and temporal context. *Percept Psychophys.* 1982;32(3):211–8.
90. Overvliet KE, Sayim B. Perceptual grouping determines haptic contextual modulation. *Vision Res.* 2016 Sep 1;126:52–8.
91. Levi DM. Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Res.* 2008;48(5):635–54.
92. Levi DM, Toet A, Tripathy SP, Kooi FL. The effect of similarity and duration on spatial interaction in peripheral vision. *Spat Vis.* 1994;8(2):255–79.
93. Chung ST, Levi DM, Legge GE. Spatial-frequency and contrast properties of crowding. *Vision Res.* 2001;41(14):1833–50.
94. Nazir TA. Effects of lateral masking and spatial precueing on gap-resolution in central and peripheral vision. *Vision Res.* 1992;32(4):771–7.
95. Bouma H. Interaction effects in parafoveal letter recognition. *Nature.* 1970;226(5241):177–8.
96. Toet A, Levi DM. The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Res.* 1992;32(7):1349–57.
97. Banks WP, Larson DW, Prinzmetal W. Asymmetry of visual interference. *Percept Psychophys.* 1979;25(6):447–56.
98. Petrov Y, Popple AV, McKee SP. Crowding and surround suppression: Not to be confused. *J Vis.* 2007;7(2):12–12.
99. He S, Cavanagh P, Intriligator J. Attentional resolution and the locus of visual awareness. *Nature.* 1996;383(6598):334–7.
100. Freeman J, Simoncelli EP. Metamers of the ventral stream. *Nat Neurosci.* 2011;14(9):1195–201.
101. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis.* 2009;9(12):13–13.
102. 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.
103. Pelli DG, Palomares M, Majaj NJ. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *J Vis.* 2004;4(12):12–12.
104. Wilkinson F, Wilson HR, Ellemberg D. Lateral interactions in peripherally viewed texture arrays. *Josa A.* 1997;14(9):2057–68.
105. Greenwood JA, Bex PJ, Dakin SC. Positional averaging explains crowding with letter-like stimuli. *Proc Natl Acad Sci.* 2009;106(31):13130–5.
106. Nandy AS, Tjan BS. Saccade-confounded image statistics explain visual crowding. *Nat Neurosci.* 2012 Mar;15(3):463–9.

107. 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.
108. Kalpadakis-Smith AV, Goffaux V, Greenwood JA. Crowding for faces is determined by visual (not holistic) similarity: Evidence from judgements of eye position. *Sci Rep*. 2018;8(1):1–14.
109. Strasburger H, Malania M. Source confusion is a major cause of crowding. *J Vis*. 2013;13(1):24–24.
110. Chung ST, Legge GE. Precision of position signals for letters. *Vision Res*. 2009;49(15):1948–60.
111. Strasburger H. Unfocussed spatial attention underlies the crowding effect in indirect form vision. *J Vis*. 2005;5(11):8–8.
112. Arman AC, Chung ST, Tjan BS. Neural correlates of letter crowding in the periphery. *J Vis*. 2006;6(6):804.
113. Beard BL, Levi DM, Klein SA. Vernier acuity with non-simultaneous targets: The cortical magnification factor estimated by psychophysics. *Vision Res*. 1997;37(3):325–46.
114. Levi DM, Klein SA, Aitsebaomo AP. Vernier acuity, crowding and cortical magnification. *Vision Res*. 1985 Jan 1;25(7):963–77.
115. Rovamo J, Virsu V. An estimation and application of the human cortical magnification factor. *Exp Brain Res*. 1979;37(3):495–510.
116. Daniel PM, Whitteridge D. The representation of the visual field on the cerebral cortex in monkeys. *J Physiol*. 1961;159(2):203.
117. Motter BC. Crowding and object integration within the receptive field of V4 neurons. *J Vis*. 2002;2(7):274–274.
118. Pinon MC, Gattass R, Sousa AP. Area V4 in Cebus monkey: extent and visuotopic organization. *Cereb Cortex N Y NY* 1991. 1998;8(8):685–701.
119. Ferrera VP, Nealey TA, Maunsell JH. Mixed parvocellular and magnocellular geniculate signals in visual area V4. *Nature*. 1992;358(6389):756–8.
120. Logothetis NK, Charles ER. The minimum motion technique applied to determine isoluminance in psychophysical experiments with monkeys. *Vision Res*. 1990;30(6):829–38.
121. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. *J Vis*. 2013;13(13):10–10.
122. Malania M, Herzog MH, Westheimer G. Grouping of contextual elements that affect vernier thresholds. *J Vis*. 2007;7(2):1–1.
123. Levi DM, Carney T. Crowding in peripheral vision: Why bigger is better. *Curr Biol*. 2009;19(23):1988–93.
124. Saarela TP, Sayim B, Westheimer G, Herzog MH. Global stimulus configuration modulates crowding. *J Vis*. 2009;9(2):5–5.
125. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis*. 2016;16(3):35–35.
126. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis*. 2012;12(10):13–13.
127. 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.
128. Kong G, Alais D, Van der Burg E. Competing distractors facilitate visual search in heterogeneous displays. *PloS One*. 2016;11(8):e0160914.
129. Van de Weijert M, Van der Burg E, Donk M. Attentional guidance varies with display density. *Vision Res*. 2019;164:1–11.
130. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. *J Exp Psychol Hum Percept Perform*. 2017;43(4):690.
131. 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.
132. Manassi M, Whitney D. Multi-level crowding and the paradox of object recognition in clutter. *Curr Biol*. 2018;28(3):R127–33.

133. Whitney D, Levi DM. Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends Cogn Sci.* 2011;15(4):160–8.
134. Gong M, Xuan Y, Smart LJ, Olzak LA. The extraction of natural scene gist in visual crowding. *Sci Rep.* 2018;8(1):1–13.
135. Louie EG, Bressler DW, Whitney D. Holistic crowding: Selective interference between configural representations of faces in crowded scenes. *J Vis.* 2007;7(2):24–24.
136. Farzin F, Rivera SM, Whitney D. Holistic crowding of Mooney faces. *J Vis.* 2009;9(6):18–18.
137. Livne T, Sagi D. Configuration influence on crowding. *J Vis.* 2007;7(2):4–4.
138. Banks WP, White H. Lateral interference and perceptual grouping in visual detection. *Percept Psychophys.* 1984;36(3):285–95.
139. Intriligator J, Cavanagh P. The spatial resolution of visual attention. *Cognit Psychol.* 2001;43(3):171–216.
140. Chakravarthi R, Cavanagh P. Temporal properties of the polarity advantage effect in crowding. *J Vis.* 2007;7(2):11–11.
141. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychol Rev.* 2017;124(4):483.
142. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. *J Vis.* 2019;19(7):15–15.
143. Manassi M, Hermens F, Francis G, Herzog MH. Release of crowding by pattern completion. *J Vis.* 2015;15(8):16–16.
144. Volokitin A, Roig G, Poggio TA. Do deep neural networks suffer from crowding? In: *Advances in Neural Information Processing Systems*. 2017. p. 5628–38.
145. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Prepr ArXiv181112231.* 2018;
146. Holland JH. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* MIT press; 1992.

Chapter 1: Beyond Bouma's window - How to explain global aspects of crowding?

Doerig, A.¹, Bornet, A.¹, Rosenholtz, R.², Francis, G.³, Clarke, A.M.⁴, Herzog, M. H.¹

¹Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne, Switzerland

²Department of Brain and Cognitive Sciences, Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

³Department of Psychological Sciences, Purdue University, USA

⁴Laboratory of Computational Vision, Psychology Department, Bilkent University, Ankara, Turkey

Abstract

In crowding, perception of an object deteriorates in the presence of nearby elements. Although crowding is a ubiquitous phenomenon, since elements are rarely seen in isolation, to date there exists no consensus on how to model it. Previous experiments showed that the global configuration of the entire stimulus must be taken into account. These findings rule out simple pooling or substitution models and favor models sensitive to global spatial aspects. In order to investigate how to incorporate global aspects into models, we tested a large number of models with a database of forty stimuli tailored for the global aspects of crowding. Our results show that incorporating grouping like components strongly improves model performance.

Author Summary

Visual crowding highlights interactions between elements in the visual field. For example, an object is more difficult to recognize if it is presented in clutter. Crowding is one of the most fundamental aspects of vision, playing crucial roles in object recognition, reading and visual perception in general, and is therefore an essential tool to understand how the visual system encodes information based on its retinal input. Classic models of crowding have focused only on local interactions between neighboring visual elements. However, abundant experimental evidence argues against local processing, suggesting that the global configuration of visual elements strongly modulates crowding. Here, we tested all available models of crowding that are able to capture global processing across the entire visual field. We tested 12 models including the Texture Tiling Model, a Deep Convolutional Neural Network and the LAMINART neural network with large scale computer simulations. We found that models incorporating a grouping component are best suited to explain the data. Our results suggest that in order to understand vision in general, mid-level, contextual processing is inevitable.

Introduction

When an element is presented in the presence of nearby elements or clutter, it becomes harder to perceive, a well-known effect called crowding. One of the main characteristics of crowding is that the element itself is not invisible, contrary to contrast- and backward-masking; rather its features appear jumbled and distorted (Fig 1). Crowding is a ubiquitous phenomenon because elements are rarely encountered in isolation in everyday situations (Fig 1c). Thus, understanding crowding is crucial for understanding vision in general. For about half century, the consensus was that flankers interfere with a target element only when placed within a spatially restricted window around the target, the so-called Bouma law (Fig 1b; 1–4):

$$\text{Size of Bouma's window} \approx 0.5 * \text{eccentricity}$$

Classic models of crowding proposed that early visual areas, such as V1, process the features of stimuli with high precision. Crowding occurs when neural signals are pooled along the visual hierarchy, e.g., when V2 neurons pool neural signals from V1 neurons (5). Hence, in line with classic hierarchical feedforward processing (Fig 2a), crowding may be seen as a natural consequence of object recognition in the visual system. For example, a hypothetical neuron coding for a square might respond to signals from neurons coding for the lines making up the square. In order to achieve translational invariance, the square neuron is sensitive to lines all over its receptive field and pools this information in order to decide whether a square is present. According to this logic, crowding occurs when elements that do not belong to the same object are pooled. In this sense, crowding is an unwanted by-product of object recognition and, for this reason, a bottleneck of vision (for a review, see 2,6). Other models have proposed that performance in crowding deteriorates because features of the target are substituted for features of the flanking elements. As mentioned, all these models are local in the sense that crowding is determined by nearby elements only. Based on these two lines of thought, pooling and substitution, researchers have suggested that with more flankers, performance deteriorates because more irrelevant features are pooled or substituted.



Fig 1. Crowding. **a.** In crowding, the perception of a target element deteriorates in the presence of nearby elements. When fixating the left cross, the target letter V on the right is hard to identify because of the nearby flankers. **b.** The task is easier than in (a), because the flankers are further away from the target letter V. Bouma's law states that crowding occurs only when flankers are sufficiently close to the target, within the so-called Bouma's window. **c.** Crowding is a ubiquitous phenomenon since elements are rarely seen in isolation. For example, when fixating the central red dot, the child on the left is easier to detect because it is not surrounded by nearby flankers, as is the child on the right.

The understanding of crowding has largely changed in the last decade. For example, it has been shown that detailed information can survive crowding (7,8). Crowding occurs in the fovea and is not restricted to the periphery, contrary to earlier proposals (9,10). Most importantly for the present discussion, performance depends on elements far outside of Bouma's window. For example, in supercrowding, elements outside of Bouma's window decrease performance beyond the decrement arising from elements within the window (11). Surprisingly, adding flankers can even *reduce* crowding, and such *uncrowding* effects can depend on elements outside of Bouma's window (Fig 2; 9,12–16, review: 17). For example, observers performed a vernier discrimination task. When a surrounding square was added to the vernier, the task became much more difficult: a classic crowding effect. However, adding more flanking squares *improved* performance gradually, i.e., performance improved the more squares were presented (18; Fig 2b). The entire line of squares extends over 17 degrees in the right visual field, while the single vernier offset threshold is less than 200'' (Fig 2d). Hence, performance is not exclusively determined by local interactions: fine-grained vernier acuity in the range of about 200'' depends on elements as far away as 8.5 degrees - a ratio of two orders of magnitude, extending far beyond Bouma's window. Moreover, performance depends on the overall configuration (19). For example, in three-by-seven displays of squares and stars (Fig 2c), a shift of the central row changes performance strongly (Fig 2c, 4th and 5th configurations).

Similar effects were found with stimuli other than verniers (20,21), as well as in auditory (22) and haptic crowding (23).

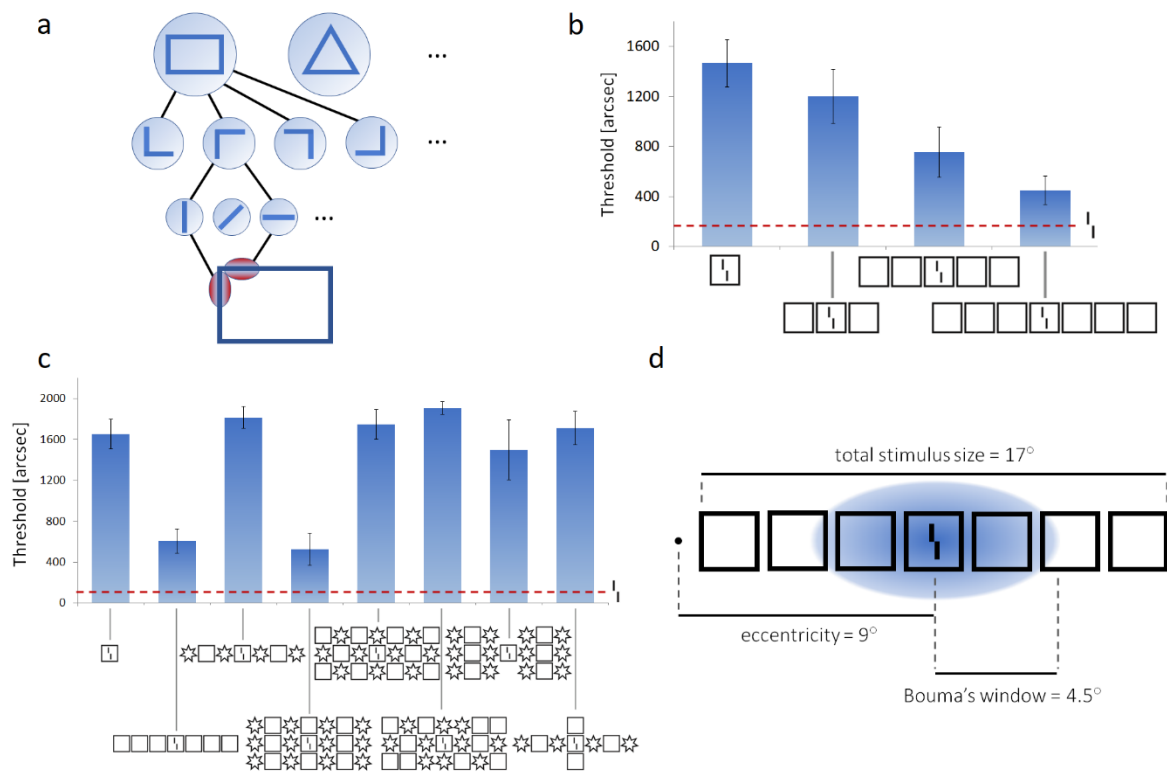


Fig 2. a. Standard view of visual processing. First, edges are detected by low-level neurons with small receptive fields. Higher level neurons pool signals from lower level neurons in a hierarchical, feedforward manner, creating higher level representations of objects by combining low-level features (24,25). For example, two low-level edge detectors may be combined to create a “corner” representation. Four such corner detectors can be assembled to create a rectangle representation. Receptive field size naturally increases along this pathway since, for example, a rectangle covers larger parts of the visual field than the lines making up the rectangle. **b. Uncrowding.** Observers performed a vernier discrimination task. The y-axis shows the threshold, for which observers correctly discriminate the vernier offset in 75% of trials (performance is good when the threshold is low). First, only a vernier is presented, an easy task (performance for this condition is shown as the dashed horizontal line). Then, a flanking square is added making the task much more difficult (a). This is a classic crowding effect. Importantly, adding more flanking squares improved performance gradually, i.e., performance improved the more squares are presented (18). We call this effect *uncrowding*. **c. The global configuration** of the entire stimulus determines crowding. Performance is strongly affected by elements far away from the target as shown in these examples (14). **d. Performance is not determined by local interactions only.** In this display, fine-grained vernier acuity of about 200'' depends on elements as far away as 8.5 degrees - a difference of two orders of magnitude, extending far beyond Bouma's window.

Because they cannot produce long-range effects, local models cannot explain the global aspects of crowding. Here, we tested which global models, integrating information across large parts of the visual field, can explain global effects on crowding (see Fig 3 for a list). We also tested the most prominent local models to verify our hypothesis that local models are inadequate to explain global aspects of crowding. The models are described in detail in the supplementary information (see [Suppl. Inf. A](#)).

The models differ with respect to four criteria:

Spatial extent: Local vs. Global. In a *local* model, elements far from the target do not exert any effects on the target. By contrast, in a *global* model, any element in the visual field may potentially interfere with target processing.

Mechanism of interference: Pooling, substitution, or other?

Organisation: Feed-forward (features at a given level are only affected by lower level features) vs. recurrent processing (features at a given level can be affected by lower or higher-level features).

Grouping component: Does the model incorporate a grouping component? Certain models explicitly compute grouping-like aspects by determining which low-level elements should belong to the same higher-level group. Only elements within a group interfere with each other.

	Spatial extent		Mechanism			Organisation		Grouping
	Local	Global	Pooling	Substitution	Other	Feed-forward	Recurrent	
Classic pooling	✓		✓			✓		
Classic substitution	✓			✓		✓		
Population coding	✓				✓	✓		
Epitomes	✓			✓		✓		
Single Texture Model		✓	✓			✓		
Texture Tiling	✓		✓			✓		
Deep Textures		✓	✓			✓		
Saccade-confounded Statistics		✓	✓			✓		
Wilson & Cowan-type net		✓	✓				✓	
Fourier Analysis		✓			✓	✓		
Zhaoping's V1 recurrent model		✓	✓				✓	
LAMINART		✓	✓				✓	✓
Hierarchical Sparse Selection	✓		✓				✓	
Alexnet (Convolutional neural network)	✓		✓			✓		

Fig 3. The models tested and their characteristics. Models may integrate information locally or globally, and the interference mechanism may be pooling, substitution, or other. Models are feed-forward or recurrent and may or may not compute grouping-like aspects of the stimulus. The aim of the current work is to investigate which models can explain the global effects of crowding.

Methods

To test the models, we used human data from previous work exploring the crowding/uncrowding phenomena (9,10,14,16,18,19). The stimulus database comprises 40 different stimuli belonging to 12 different categories: circles, Gestalts, hexagons, irregular1, irregular2, lines, octagons, patternIrregular, patternStars, squares and stars. An example of each category is shown in Fig 4. Behavioral results can be found in the original papers. In each category, we have the vernier target alone, plus crowding and uncrowding configurations. All the stimuli are shown in Fig 5 and behavioural results can be found in the original papers. With a few exceptions (see details in [Suppl. Inf. A](#)), we ran each model on all stimuli. For some models, we could not use the entire database because computation time was too long (deep convolutional networks, LAMINART, Texture Tiling Model), or because the model was not adapted to accommodate certain kinds of stimuli (Population Coding). Human and model results are summarized in the Results section (Fig 5-6). The code we used is available online at <https://github.com/adriendoerig/beyond-boumas-window-code> (except the Texture Tiling Model, which Rosenholtz and colleagues will share in a forthcoming publication). All the results can be found at <https://github.com/adriendoerig/beyond-boumas-window-results>.

There are two fundamentally different approaches to measure model performance. First, a linking hypothesis may be used to relate model output to performance (both are scalar numbers). For example, template matching computes how similar the model output is to the target image. If they are similar, performance is good. The second, textural approach is used to quantify performance in textural models. The idea is that peripheral vision is ambiguous because information is compressed by summary statistics. If a model uses a proper algorithm for representing these ambiguities, presenting the processed image in the fovea should lead to similar human performance as presenting the original unprocessed image in the periphery (26). Accordingly, to measure the performance of textural algorithms, the stimuli are fed through a texture synthesis procedure. Then, observers freely examine the output image and report vernier orientation. If this task is easy, performance is good. For each model, we used the linking hypothesis proposed by the original authors when available. When this was not possible (for example for Alexnet, which has never been applied to crowding results before), we detail which linking hypothesis we used in the corresponding section. In the Supplementary

information, we present, first, textural models (Suppl. Inf. A; SA1-SA4) and, second, models using a linking hypothesis (Suppl. Inf. A; SA5-SA12).

An important point is that different readouts lead to different results. Hence, the different methods of model evaluation used here could affect our results. However, we are mainly interested in qualitative rather than quantitative comparisons and the readout functions we used cannot confuse crowding and uncrowding. More specifically, the readout processes we use produce results monotonically linked to the model outputs. Hence, they cannot confuse uncrowding cases (a U-shape function where the vernier alone condition leads to good performance, a single flanker deteriorates performance, and multiple flankers lead again to good performance) with cases that do not show uncrowding (a monotonic function where the vernier alone condition leads to good performance, a single flanker deteriorates performance, and multiple flankers deteriorate performance even more).

Because different models were evaluated differently, it was impossible to come up with one performance measure and to compare models via something like the Akaike Information Criterion. However, despite this variety of performance measures, our results are qualitatively unambiguous: each model either is capable of producing uncrowding, or it is not. We took the parameters directly from the original models whenever possible. Otherwise, we tried our best to search the parameter space. We cannot exclude that other combinations of parameters fit the dataset better. However, we will argue that models that cannot produce uncrowding fail to do so for principled reasons, and not because of poor parameter choices (see Discussion).

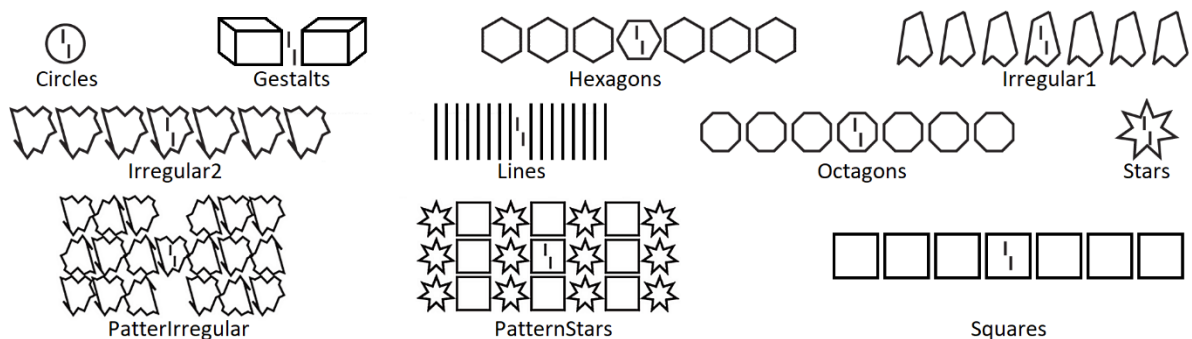


Fig 4. Stimulus categories. We used 40 different stimuli from 11 different categories. The task was always to report the offset direction of the central vernier. This figure shows one example from each category. The stimulus database is tailored to test for global effects such as uncrowding. Human data was taken from previous work (9,10,14,16,18,19). Human and model results are summarized in the discussion (Fig 5 shows the results for all stimuli and models).

Results

















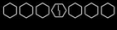

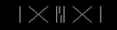











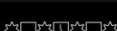




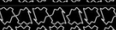
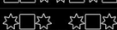
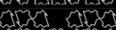


Crowded conditions		Epitomes	Single Texture	Texture Tiling	Wilson & Cowan	Fourier Analysis	HSS	Zhaoqing's V1 model	LAMINART	AlexNet (CNN)	Uncrowded conditions		Epitomes	Single Texture	Texture Tiling	Wilson & Cowan	Fourier Analysis	HSS	Zhaoqing's V1 model	LAMINART	AlexNet (CNN)
1 square ⁽³⁾											7 squares ⁽³⁾										
1 circle ⁽⁴⁾											7 circles ⁽⁴⁾										
1 irreg1 ⁽⁴⁾											7 irreg1 ⁽⁴⁾										
1 irreg2 ⁽⁴⁾											7 irreg2 ⁽⁴⁾										
1 4-star ⁽⁴⁾											7 4-star ⁽⁴⁾										
1 7-star ⁽⁴⁾											7 7-star ⁽⁴⁾										
1 octagon ⁽⁴⁾											7 octagons ⁽⁴⁾										
1 hexagon horizontal ⁽⁴⁾											7 hexagons horizontal ⁽⁴⁾										
hexagons G ⁽⁴⁾											gestalts boxes ⁽²⁾										
gestalts cross ⁽²⁾											gestalts boxes and crosses ⁽²⁾										
gestalts scrambled cuboids ⁽¹⁾											gestalts cuboids ⁽¹⁾										
2 lines short ⁽²⁾											16 lines short ⁽²⁾										
2 lines long ⁽²⁾											16 lines long ⁽²⁾										
2 lines equal ⁽²⁾											pat stars D ⁽⁴⁾										
16 lines equal ⁽²⁾											pat irreg C ⁽⁴⁾										
pat stars C ⁽⁴⁾											pat irreg D ⁽⁴⁾										
pat stars E ⁽⁴⁾											pat irreg E ⁽⁴⁾										
pat stars F ⁽⁴⁾											pat irreg F ⁽⁴⁾										
pat stars G ⁽⁴⁾											pat irreg G ⁽⁴⁾										
pat stars H ⁽⁴⁾											pat irreg H ⁽⁴⁾										

Fig 5. Summary of results. Detailed results for each model can be found in the Supplementary Information (Suppl. Inf. A). Here, the results for all models (columns) are summarized. In black, the left panel displays all crowding stimuli and the right panel displays all uncrowding stimuli (i.e., better performance when extra elements are added to the crowded condition) as observed in human data (rows). Superscript numbers indicate which publication the results are taken from (1: Sayim, Westheimer & Herzog (16); 2: Manassi et al. (10); 3: Manassi, Sayim & Herzog (18); 4: Manassi et al. (14)). Red indicates that the model predicts crowding, green indicates uncrowding and gray indicates that we did not run the model on the stimulus. Only the LAMINART model is capable of producing uncrowding consistently. Fourier and the Wilson-Cowan network produce uncrowding but

suffer from overfitting (see Discussion). For these two models, we provide the results for the best parameters. For example, the Wilson and Cowan network with different parameters can explain the lines category but then it cannot explain the squares categories.

Model	Captures effects of		Grouping component
	Crowding	Uncrowding	
Local Substitution (Strasburger et al., 1991; Ester et al., 2014)	Yes	No	No
Local pooling (Parks et al., 2001; Pelli et al., 2004)	Yes	No	No
Epitomes (Jojic et al., 2003)	Yes	No	No
Single Texture Model (Portilla & Simoncelli, 2000)	Yes	No	No
Texture Tiling Model (Freeman & Simoncelli, 2011; Rosenholtz et al., 2012)	Yes	No	No
Deep Textures (Gatys, Ecker & Bethge, 2015)	Yes	No	No
Wilson & Cowan Network (Wilson & Cowan, 1973; Clarke, unpublished)	Yes	Yes	No
Zhaoping's V1 recurrent model (Zhaoping, 1999)	Yes	No	No
LAMINART (Francis et al., 2017)	Yes	Yes	Yes
AlexNet (Convolutional Neural Network)	Yes	No	No
Hierarchical Sparse Selection (Chaney, Fischer & Whitney, 2014)	Yes	No	No
Saccade Confounded Summary Statistics (Nandy & Tjan, 2012)	Yes	No	No
Population coding (van den Berg et al., 2010, Dayan & Solomon, 2010, Harrison & Bex, 2015)	Yes	No	No
Fourier model (Manassi et al., 2015)	Yes	Yes	No

Fig 6. All models produce crowding, but only the Fourier, Wilson and Cowan and LAMINART models produce uncrowding. The Fourier and the Wilson and Cowan model overfit and thus do not capture general principles (see Fig 5). The LAMINART is the only model, which explicitly computes grouping like aspects (illusory contours, see [Suppl. Inf. A; SA7](#)) and segments the image into different layers.

Discussion

For decades, crowding was thought to be fully determined by nearby elements. For this reason, target elements were presented only with a few nearby elements, and models were local in nature. However, experiments of the last two decades have shown that elements far beyond Bouma's window can strongly affect performance. Crowding can become stronger (11) or weaker (9,12–15) when elements are presented outside Bouma's window. Hence, local models cannot provide a complete account of crowding. In addition to spatial extent, it is the specific stimulus configuration that determines crowding. Configurational effects are not small modulations of crowding but have large effect sizes and, more importantly, these effects can *qualitatively* change the pattern of results. For example, in Fig 2b, performance changes in a non-linear U-shaped fashion with best performance for the unflanked target, strong crowding for few flankers, and weaker crowding when flankers make up a regular configuration.

A major question is at which computational level crowding occurs. In local models, only nearby elements interfere with target processing, often due to low level mechanisms such as pooling. In global models, features across the entire visual field are potentially important. Global interactions may be restricted to low level features, such as the orientations of the stimulus elements. At the other extreme, explicitly computing objects may turn out to be necessary (for example the squares in Fig 2). Likewise, face crowding may or may not necessitate the explicit computation of faces (7,50,57,58). For this reason, other global models explicitly compute grouping-like aspects. Only elements within a group interfere with each other. Classically, models restricting themselves to lower level features are given priority because they offer more parsimonious explanations.

Model comparison

Here, we investigated all available models suited to explain the global aspects of crowding. All models (leaving aside Deep Textures, which was never proposed to explain crowding with laboratory stimuli) produced crowding comparable to the human data. However, only the LAMINART model was consistently able to produce *uncrowding* (see Fig 5, more details in [Suppl. Inf. A; SA7](#)). The Wilson and Cowan network (more details in [Suppl. Inf. A; SA5](#)) produced uncrowding only for the squares category (and to a lesser extent for the lines and irregular1 when they were used as training sets). The Fourier model (more details in [Suppl. Inf. A; SA12](#))

produced uncrowding only for the Gestalts and lines stimuli. In both models, uncrowding depended heavily on parameter values, a signature of overfitting. In the Wilson and Cowan network, the end-stopped receptive fields grouped elements of similar size, but this did not generalize to explain other global effects.

We think there are principled reasons why most models cannot reproduce most of the global uncrowding findings. First, the effects of global configuration (Fig 2b) operate on a much higher level than most models can capture. To phrase it this way, we think that human performance is based on global configurations and not on simple hidden sub-regularities, such as repeating patterns or simple summary statistics. Second, as Wallis et al. (29) put it: "Based on our experiments we speculate that the concept of summary statistics cannot fully account for peripheral scene appearance. Pooling in fixed regions will either discard (long-range) structure that should be preserved or preserve (local) structure that could be discarded. Rather, we believe that the size of pooling regions needs to depend on image content". For this reason, we think that performance in crowding cannot be explained simply as a by-product of *basic* spatial processing, e.g., by summary statistics. In contrast, which elements interfere seems to depend on the global stimulus layout. We propose that the LAMINART model can consistently produce uncrowding because it can deal with this requirement by incorporating a grouping-like process: elements linked by illusory contours are grouped together and segmented from elements in other groups. Interference happens only between elements within a group.

Another way to approach the importance of grouping for crowding is that it provides extra information that makes one condition inherently easier than another. Vernier acuity tasks are often thought to be mediated by the responses of one or more feature detector. Each feature detector might itself look like a vernier offset or might be similar to an orientation detector like a Gabor. Regardless, correct performance at the vernier task requires precise placement of the detector; a slightly misplaced detector can easily give the wrong answer, particularly when the vernier is flanked by other stimuli. Crowding induces location uncertainty. Any information that helps placing the detector – essentially any cue to the right position – would improve performance. Strong stimulus grouping could be one such cue (30). In this case too, it is crucial to understand how the brain groups visual elements across the entire visual field.

The LAMINART model links elements by illusory contours, which is a rather basic grouping mechanism (details in [Suppl. Inf. A; SA7](#)). It remains an open question whether more complex features are necessary to explain crowding/uncrowding such as an explicit computation of objects, e.g. squares, faces etc. For example, can the irregular shapes category be explained with simple contour integration? Likewise, it remains an open question whether face crowding can be explained without the explicit computation of faces.

In the LAMINART model, the grouping and interference processes are separate. Alternatively, grouping and interference may be intimately linked. One possibility is that the groups correspond to optimal statistical representations. For example, elements may form a group when they can be well compressed by summary statistics. In this scenario, grouping is part of the summary statistics process itself. There are probably many other ways in which grouping may play a role.

A major problem with the grouping approach is the lack of a well-defined, objective measure of grouping. If there is no objective measure, groups can be chosen ad hoc to explain experimental results, leading to circular explanations. As a first step towards an objective measure of grouping, subjective measures (i.e., asking observers to report what they feel belongs to a group) can complement studies. Such subjective ratings about perceptual groups have correlated well with psychophysical performance levels (10).

Future Models

As we have shown, none of the current models can fully explain (un)crowding. What would the model of the future look like? What components are crucial?

First, as mentioned earlier, we can rule out local models because elements across large parts of the visual field influence perception of the target.

Second, to explain the complex effects of spatial configurations in crowding, our results suggest that grouping-like, mid or higher-level aspects need to be incorporated in a model. However, the exact nature of this process is unknown. For example, it may or may not be that mid-level processing is sufficient. In addition, the incorporation of higher-level processes does not exclude the additional use of summary statistics and other lower level components. The grouping stage is difficult to study because of the seemingly infinite number of possible visual

configurations. We believe that new tools are needed to help navigate the huge search space effectively. For example, Van der Burg, Olivers, & Cass (31) have proposed a genetic algorithm to find configurational features important for crowding.

Third, we cannot rule out feedforward models. Indeed, it is a mathematical fact that any recurrent model can be “unfolded” into a feed-forward network (32–34). However, these feedforward models are usually extremely large and computationally expensive. For this reason, we suggest that models with feedback connections are much more likely to be able to explain how complex spatial configurations influence target processing. For example, higher level grouping processing, such as computing the squares and grouping them together, may feed back to lower level processing of the target, i.e., the vernier. Support for this hypothesis comes from the finding that the deep neural networks could not produce uncrowding, presumably because high-level features cannot influence low-level processing.

Fourth, the nature of interference remains unclear. One option is that interference occurs *during* complex spatial processing by an unknown mechanism. Another option is that the classic interference mechanisms operate *after* complex spatial processing is accomplished. For example, pooling may occur only for grouped elements. In the same line of reasoning, Chaney et al. (35), Van den Berg et al. (36) and Harrison & Bex (37) noted that adding a grouping stage to their interference mechanism may help explain a wider range of results. Combining complex spatial processing with good interference mechanisms may, therefore, allow for a happy marriage between interference- and grouping-based mechanisms leading to a truly unified model of crowding.

Conclusion

The global stimulus configuration plays a crucial role in crowding, which cannot be captured by local models. For this reason, we propose that models of crowding need to include grouping like processes. While our results show that none of the current models lacking a grouping process can explain the global uncrowding phenomena, they may be good candidates for a potential second, interference stage.

How are basic features of the visual field grouped to form objects? The most successful model we analyzed, the LAMINART variation, suggests that this is done by linking features together by illusory contours. Further work is needed to assess how far this mechanism can go and what additional components are necessary, such as summary statistics. For example, the groups may correspond to optimal statistical representations (elements that can easily be compressed using summary statistics would form a group).

Most importantly, large scale, configurational effects are not restricted to visual crowding with vernier targets. Uncrowding occurs also for letters and Gabors (38), as well as in audition (22) and haptics (23). Similar effects are found in backward masking (39) and overlay masking (16,40). Hence, crowding is only a special case of contextual processing. Vision research has largely missed these aspects because of the use of well-controlled stimuli, which are usually presented in isolation or, as in crowding, with only a few nearby flankers. Our results suggest that in order to understand vision in general, a mid-level, contextual processing stage is inevitable.

References

1. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Res.* 1973 Apr 1;13(4):767–82.
2. Levi DM. Crowding-An essential bottleneck for object recognition: A mini-review. *Vision Res.* 2008;48(5):635–54.
3. Pelli DG, Palomares M, Majaj NJ. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *J Vis.* 2004 Dec 1;4(12):12–12.
4. Strasburger H, Harvey LO, Rentschler I. Contrast thresholds for identification of numeric characters in direct and eccentric view. *Percept Psychophys.* 1991 Nov 1;49(6):495–508.
5. Pelli DG. Crowding: a cortical constraint on object recognition. *Curr Opin Neurobiol.* 2008 Aug 1;18(4):445–51.
6. Pelli DG, Tillman KA. The uncrowded window of object recognition. *Nat Neurosci.* 2008;11(10):1129–35.
7. Manassi M, Whitney D. Multi-level Crowding and the Paradox of Object Recognition in Clutter. *Curr Biol.* 2018 Feb 5;28(3):R127–33.
8. Whitney D, Haberman J, Sweeny TD. From textures to crowds: multiple levels of summary statistical perception. *New Vis Neurosci.* 2014;695–710.
9. Malania M, Herzog MH, Westheimer G. Grouping of contextual elements that affect vernier thresholds. *J Vis.* 2007 Jan 2;7(2):1–1.
10. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis.* 2012 Sep 1;12(10):13–13.
11. Vickery TJ, Shim WM, Chakravarthi R, Jiang YV, Luedeman R. Supercrowding: Weakly masking a target expands the range of crowding. *J Vis.* 2009 Feb 1;9(2):12–12.
12. Banks WP, Larson DW, Prinzmetal W. Asymmetry of visual interference. *Percept Psychophys.* 1979 Nov 1;25(6):447–56.
13. Livne T, Sagi D. Configuration influence on crowding. *J Vis.* 2007 Jan 2;7(2):4–4.
14. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis.* 2016 Feb 1;16(3):35–35.
15. Pöder E. Crowding, feature integration, and two kinds of “attention.” *J Vis.* 2006 Feb 1;6(2):7–7.
16. Sayim B, Westheimer G, Herzog MH. Gestalt factors modulate basic spatial vision. *Psychol Sci.* 2010;21(5):641–4.
17. Herzog MH, Sayim B, Chicherov V, Manassi M. Crowding, grouping, and object recognition: A matter of appearance. *J Vis.* 2015 May 1;15(6):5–5.
18. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. *J Vis.* 2013 Nov 1;13(13):10–10.
19. Herzog MH, Manassi M. Uncorking the bottleneck of crowding: a fresh look at object recognition. *Curr Opin Behav Sci.* 2015 Feb;1:86–93.
20. Chakravarthi R, Pelli DG. The same binding in contour integration and crowding. *J Vis.* 2011 Jul 5;11(8):10–10.
21. Livne T, Sagi D. Multiple levels of orientation anisotropy in crowding with Gabor flankers. *J Vis.* 2011 Nov 1;11(13):18–18.
22. Oberfeld D, Stahn P. Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PloS One.* 2012;7(10):e48054.
23. Overvliet KE, Sayim B. Perceptual grouping determines haptic contextual modulation. *Vision Res.* 2016 Sep 1;126(Supplement C):52–8.
24. DiCarlo JJ, Zoccolan D, Rust NC. How Does the Brain Solve Visual Object Recognition? *Neuron.* 2012 Feb 9;73(3):415–34.
25. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci.* 1999;2(11).
26. Fischer J, Whitney D. Object-level visual information gets through the bottleneck of crowding. *J Neurophysiol.* 2011;106(3):1389–98.

27. Kalpadakis-Smith A, Goffaux V, Greenwood J. Crowding for faces is determined by visual (not holistic) similarity: Evidence from judgements of eye position. 2017;
28. Sun H-M, Balas B. Face features and face configurations both contribute to visual crowding. *Atten Percept Psychophys*. 2015 Feb 1;77(2):508–19.
29. Wallis T, Funke C, Ecker A, Gatys L, Wichmann F, Bethge M. Towards matching peripheral appearance for arbitrary natural images using deep features. *J Vis*. 2017;17(10):786–786.
30. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. *J Vis*. 2019;19(7):15–15.
31. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. *J Exp Psychol Hum Percept Perform*. 2017;43(4):690.
32. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw*. 1989;2(5):359–66.
33. Schäfer AM, Zimmermann HG. Recurrent Neural Networks Are Universal Approximators. In: *Artificial Neural Networks – ICANN 2006* [Internet]. Springer, Berlin, Heidelberg; 2006 [cited 2017 Dec 5]. p. 632–40. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/11840817_66
34. Werbos PJ. Generalization of backpropagation with application to a recurrent gas market model. *Neural Netw*. 1988 Jan 1;1(4):339–56.
35. Chaney W, Fischer J, Whitney D. The hierarchical sparse selection model of visual crowding. *Front Integr Neurosci*. 2014;8.
36. 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.
37. Harrison WJ, Bex PJ. Reply to Pachai et al. *Curr Biol*. 2016;26(9):R353–4.
38. Saarela TP, Westheimer G, Herzog MH. The effect of spacing regularity on visual crowding. *J Vis*. 2010;10(10):17–17.
39. Herzog MH, Fahle M. Effects of grouping in contextual modulation. *Nature*. 2002 Jan;415(6870):433.
40. Saarela TP, Sayim B, Westheimer G, Herzog MH. Global stimulus configuration modulates crowding. *J Vis*. 2009;9(2):5–5.
41. Jojic N, Frey BJ, Kannan A. Epitomic analysis of appearance and shape. In: *Proceedings Ninth IEEE International Conference on Computer Vision*. 2003. p. 34–41 vol.1.
42. Portilla J, Simoncelli EP. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *Int J Comput Vis*. 2000 Oct 1;40(1):49–70.
43. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis*. 2009 Nov 1;9(12):13–13.
44. Zhang X, Huang J, Yigit-Elliott S, Rosenholtz R. Cube search, revisited. *J Vis*. 2015;15(3):9–9.
45. Freeman J, Simoncelli EP. Metamers of the ventral stream. *Nat Neurosci*. 2011 Sep;14(9):1195–201.
46. Keshvari S, Rosenholtz R. Pooling of continuous features provides a unifying account of crowding. *J Vis*. 2016 Feb 1;16(3):39–39.
47. Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L. A summary statistic representation in peripheral vision explains visual search. *J Vis*. 2012 Apr 2;12(4):14–14.
48. Gatys L, Ecker AS, Bethge M. Texture Synthesis Using Convolutional Neural Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems 28* [Internet]. Curran Associates, Inc.; 2015 [cited 2017 Oct 18]. p. 262–270. Available from: <http://papers.nips.cc/paper/5633-texture-synthesis-using-convolutional-neural-networks.pdf>
49. Wallis TSA, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *J Vis*. 2017 Oct 1;17(12):5–5.
50. Wilson HR, Cowan JD. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*. 1973 Sep 1;13(2):55–80.
51. Hermens F, Luksys G, Gerstner W, Herzog MH, Ernst U. Modeling spatial and temporal aspects of visual backward masking. *Psychol Rev*. 2008;115(1):83.

52. Panis S, Hermens F. Time course of spatial contextual interference: Event history analyses of simultaneous masking by nonoverlapping patterns. *J Exp Psychol Hum Percept Perform*. 2014;40(1):129.
53. Clarke AM, Herzog MH, Francis G. Visual crowding illustrates the inadequacy of local vs. global and feedforward vs. feedback distinctions in modeling visual perception. *Front Psychol*. 2014;5.
54. Li Z. Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Netw Comput Neural Syst*. 1999;10(2):187–212.
55. Zhaoping L. V1 mechanisms and some figure–ground and border effects. *J Physiol-Paris*. 2003;97(4):503–15.
56. Cao Y, Grossberg S. A laminar cortical model of stereopsis and 3D surface perception: closure and da Vinci stereopsis. *Spat Vis*. 2005 Nov 1;18(5):515–78.
57. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. 2017;
58. Grossberg S. Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Netw*. 2017;87:38–95.
59. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015 May 28;521(7553):436–44.
60. Lin HW, Tegmark M, Rolnick D. Why Does Deep and Cheap Learning Work So Well? *J Stat Phys*. 2017 Sep 1;168(6):1223–47.
61. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
62. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: <https://www.tensorflow.org/>
63. Nandy AS, Tjan BS. Saccade-confounded image statistics explain visual crowding. *Nat Neurosci*. 2012;15(3):463–9.
64. Harrison WJ, Bex PJ. A Unifying Model of Orientation Crowding in Peripheral Vision. *Curr Biol*. 2015 Dec 21;25(24):3213–9.
65. Dayan P, Solomon JA. Selective Bayes: Attentional load and crowding. *Vision Res*. 2010 Oct 28;50(22):2248–60.
66. Pachai MV, Doerig AC, Herzog MH. How best to unify crowding? *Curr Biol*. 2016 May 9;26(9):R352–3.
67. Agaoglu MN, Chung ST. Can (should) theories of crowding be unified? *J Vis*. 2016;16(15):10–10.

Chapter 2: Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing

Bornet, A.¹, Choung, O. H.¹, Doerig, A.^{1,2}, Whitney, D.^{3,4,5}, Herzog, M. H.¹ & Manassi M.⁶

¹Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Switzerland

²Department of Brain and Cognitive Sciences, Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

³Department of Psychology, University of California, Berkeley, CA, USA

⁴Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA

⁵Vision Science Group, University of California, Berkeley, CA, USA

⁶School of Psychology, University of Aberdeen, King's College, Aberdeen, UK

Abstract

In visual crowding, the perception of a target deteriorates in the presence of nearby flankers. Traditionally, target-flanker interactions have been considered as local, mostly deleterious, low-level and feature specific, occurring when information is pooled along the visual processing hierarchy. Recently, a vast literature of high-level effects in crowding (grouping effects and face-holistic crowding in particular) led to a completely new understanding of crowding, as a global, complex, and multi-level phenomenon that cannot be captured or explained by simple pooling models. It was recently argued that these high-level effects may still be captured by more sophisticated pooling models, such as the Texture Tiling model (TTM). Unlike simple pooling models, the high-dimensional pooling stage of the TTM preserves rich information about a crowded stimulus and, in principle, this information may be sufficient to drive high-level and global aspects of crowding. In addition, it was proposed that grouping effects in crowding may be explained by post-perceptual target cueing. Here, we extensively tested the predictions of the TTM on the results of six different studies that highlighted high-level effects in crowding. Our results show that the TTM cannot explain any of these high-level effects, and that the behavior of the model is equivalent to a simple pooling model. In addition, we show that grouping effects in crowding cannot be predicted by post-perceptual factors such as target cueing. Taken together, these results reinforce once more the idea that complex target-flanker interactions determine crowding and that crowding occurs at multiple levels of the visual hierarchy.

Introduction

In crowding, perception of a target strongly deteriorates when flanking elements are added (Pelli, 2008; Strasburger et al., 2011; Whitney & Levi, 2011). Classically, crowding was explained by pooling or bottleneck models where features of the target and nearby flankers are pooled within receptive fields of low-level neurons (Levi, 2008; Wilkinson et al., 1997). In line with this hypothesis, target-flanker interactions in crowding were characterized as (1) locally confined (Bouma's law; Bouma, 1970; Toet & Levi, 1992), (2) deleterious (Parkes et al., 2001; Wilkinson et al., 1997) and (3) low-level feature specific (Andriessen & Bouma, 1976; Chung et al., 2001; Levi et al., 1994, 2002).

Classic pooling models were seriously challenged by recent results in the last decade, and widely dismissed. First, elements beyond Bouma's window were shown to modulate crowding strength (Harrison et al., 2013; Malania et al., 2007; Manassi et al., 2012; Vickery et al., 2009). Second, it was shown that grouping determines crowding: depending on the stimulus configuration, adding flankers can reduce or increase crowding strength (Livne & Sagi, 2007, 2010; Malania et al., 2007; Saarela et al., 2010). Third, crowding was shown to occur at multiple levels along the visual hierarchy, e.g., for objects and faces (Kimchi & Pirkner, 2015; Louie et al., 2007; Sun & Balas, 2015). Taken together, target-flanker interactions in crowding are (1) global, (2) complex (i.e, crowding does not simply increase when more flankers are added), and (3) occur at multiple levels of the visual processing (reviews: Herzog et al., 2015, 2016; Herzog & Manassi, 2015; Manassi & Whitney, 2018). As a consequence, simple pooling models do not seem adequate to explain this large body of results (Doerig et al., 2019).

In response to this line of evidence, Rosenholtz et al. (2019) recently proposed that high-dimensional pooling models, e.g., the Texture Tiling Model (TTM; Rosenholtz, 2014; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, et al., 2012), can explain all these effects. In a first stage, the TTM computes V1-like responses from low-level, multi-scale and oriented feature detectors. In a second stage, the model pools these features locally to generate a large set of second-order correlations (high-dimensional pooling). Contrary to simple pooling models, the high-dimensional pooling stage preserves rich information which supports a fine-grained representation of the visual input and may, in principle, explain complex crowding effects at a later post-perceptual stage. Still, the TTM shares the

characteristics of the simpler pooling models: pooling occurs only in spatially confined regions, is restricted to low-level processing, and occurs at a single processing level. Crucially, if the TTM can predict all of the high-level effects in the recent literature, it means that target-flanker interactions are not high-level.

Here, we tested the TTM on a large body of evidence for high-level effects in crowding (Canas-Bajo & Whitney, 2020; Farzin et al., 2009; Manassi et al., 2012, 2013, 2015, 2016). First, we show that, in contrast to what Rosenholtz et al. (2019) claimed, the TTM does not reproduce any of the grouping effects in Manassi et al. (2012, 2013, 2015, 2016; section “TTM & Grouping Effects”). Second, we show that the TTM has the same limitations as simple pooling models, strictly dependent on flanker pixel density and blind to high-level configurational aspects (subsection “TTM & prediction power”). Third, Rosenholtz et al. (2019) argued that the grouping effects in crowding (Manassi et al., 2012, 2013, 2015, 2016) arise because different flanker configurations cue the target location in different ways and, thus, may modulate crowding strength in a later post-perceptual stage. We show that cueing plays no real role in Manassi et al. (2012, 2013, 2015, 2016; subsection “Grouping effects and target cueing”). Fourth, we show that holistic face processing can occur in peripheral vision despite low-level crowding, and that the TTM cannot reproduce this result because low-level information is lost irretrievably at the pooling stage of the model (section “TTM & Face Crowding”, single face discrimination task). Fifth, we show that the TTM cannot account for crowding between holistic representations of faces (Farzin et al., 2009; section “TTM & Face Crowding”, gender face discrimination task).

General Materials and Methods

Mongrel generation

To assess TTM performance, we generated mongrels for different stimuli, by using the code shared by Rosenholtz et al. (2019; <https://dspace.mit.edu/handle/1721.1/121152>). The TTM takes an image as input and outputs several images rather than a performance measure, such as accuracy. The outputted images, called mongrels, share the same pooled statistics as the original input image. The idea is that mongrels, when viewed foveally and for unlimited time, mimic the peripheral perception of the input image (Balas et al., 2009; Rosenholtz et al., 2019).

The TTM requires to set a radius for the fovea. Rosenholtz et al. (2019) suggested a value between 16 and 32 pixels. The latter value is what was used in Rosenholtz et al. (2019). As in preliminary pilots a value of 32 did not yield sufficiently strong crowding, we used a value of 16. In order to control for ceiling effects, we repeated some experiments with a radius of 32.

Stimulus images were taken from Manassi et al., (2012, 2013, 2015, 2016), Canas-Bajo & Whitney (2020), and Farzin et al. (2009). The layout of the stimuli was identical to the original publications. Every pixel was 1/30 degrees of the stimulus used in the original experiment (i.e., the resolution was 30 pixels per degree). In the original experiment of Manassi et al. (2012, 2015), stimuli were displayed on oscilloscopes. Here, we adapted our stimuli to a LCD presentation by having white lines on a black background, as in Manassi et al. (2013, 2016). All generated mongrels are available at https://github.com/albornet/TTM_Verniers_Faces_Mongrels.

Ethics

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

TTM & Grouping Effects

Methods

Stimuli

The stimuli that we used to generate the mongrels consisted of a vernier target alone or surrounded by various flanker configurations (Figure 1). The vernier target consisted of two vertical 40 arcmin lines separated by a vertical gap of 4 arcmin. The vernier target was offset either to the left or to the right. The offset size varied according to the eccentricity at which the vernier target was presented (see next paragraph).

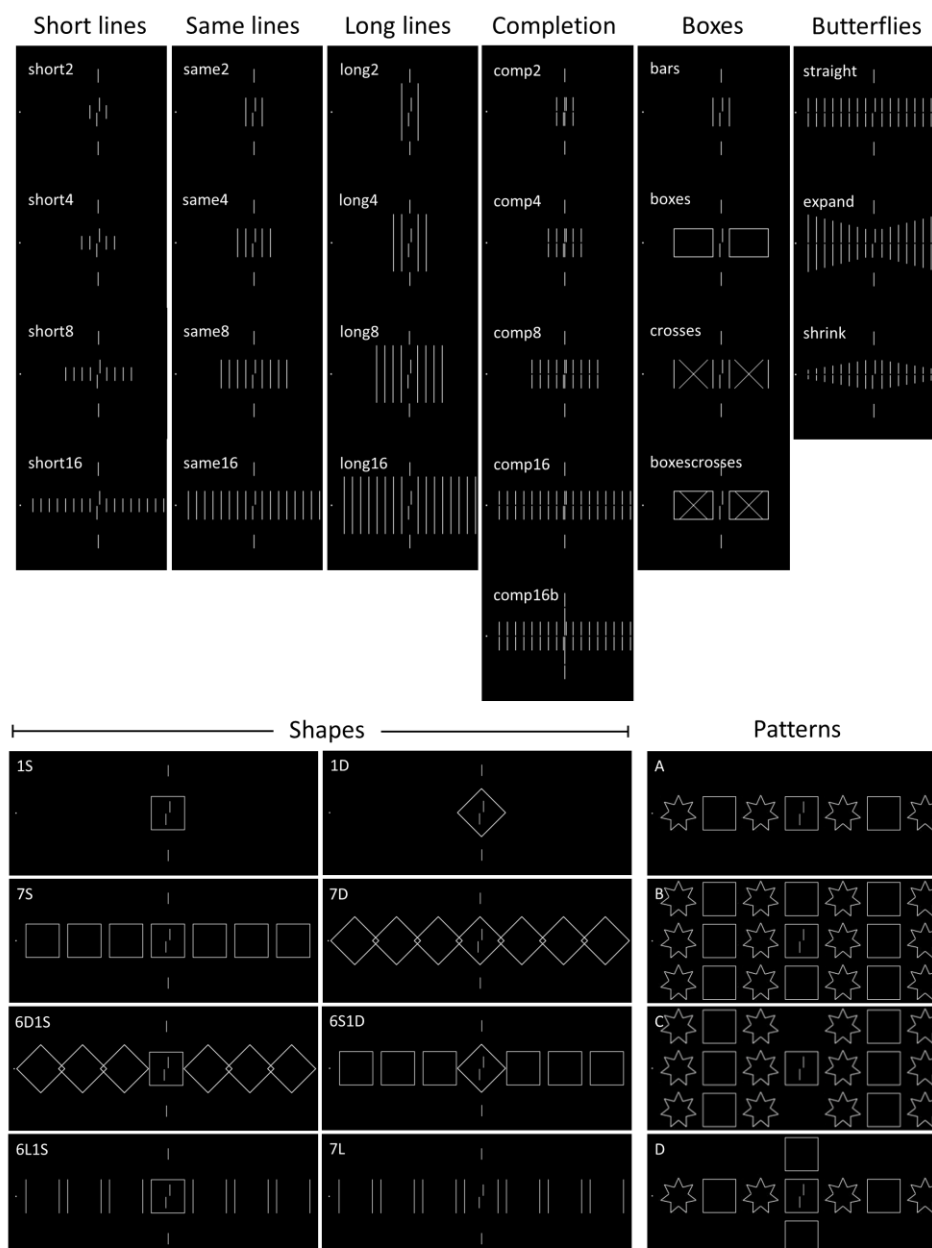


Figure 1. Stimuli used to validate the TTM. In the original experiments, observers were asked to discriminate the offset of a vernier target presented in the right hemifield and in the periphery (here shown in the center of each image), while looking at a fixation dot. Different flanker configurations were presented across the studies: “Short/Same/Long lines” and “Boxes” in Manassi et al. (2012); “Completion” and “Butterflies” in Manassi et al. (2015); “Shapes” in Manassi et al. (2013); “Patterns” in Manassi et al. (2016). In the original experiments as well as in the TTM validations, the target eccentricity was 3.88° in the “Lines”, “Boxes”, “Completion” and “Butterflies” experiments, and 9° in the “Shapes” and “Patterns” experiments.

Sixteen flanker configurations were taken from Manassi et al. (2012; Figure 1, “Short/Same/Long lines” and “Boxes”) and eight configurations from Manassi et al. (2015; Figure 1, “Completion” and “Butterflies”). For these conditions, each stimulus configuration was presented to the TTM with a vernier target eccentricity of 3.88° and a vernier offset size of 8 arcmin. Eight configurations were taken from Manassi et al. (2013; Figure 1, “Shapes”) and four configurations from Manassi et al. (2016; Figure 1, “Patterns”). For these conditions, each stimulus configuration was presented to the model with a vernier target eccentricity of 9° and a vernier offset size of 14 arcmin.

In all configurations, except the ones in the “Patterns” experiment, two vertical lines (called the “pointers”) were placed above and below the vernier target. In the original experiments, the pointers were used to reduce target location uncertainty (Manassi et al., 2012, 2013, 2015). For these configurations, we also generated mongrels using stimuli in which the pointers were removed. In total, 72 different flanker configurations were used (including the vernier alone conditions, at both eccentricities, with and without pointers). For each configuration, 30 different mongrels were generated (split equally between left and right vernier offset), for a total of 2160 unique mongrel samples shown to every participant.

Vernier offset discrimination task

Crowding strength in the TTM was quantified by performing a target discrimination task in free-viewing conditions, using the mongrels. We presented the generated mongrel images to observers and asked them to discriminate between left and right vernier offset (2AFC task). The mongrels were shown in a random order (mixed conditions).

In order to familiarize with the task, prior to the experiment, observers were shown 10 examples of the original stimulus images in which only the target was present, followed by 10 original stimulus images in which the target was embedded in different flanker configurations,

and finally 10 mongrels. In all these examples, the vernier target (or the part of the mongrel that corresponded to the vernier target) was highlighted and labelled.

13 observers performed this task (6 males, 7 females, 31.8 ± 2.9 years old). For each flanker configuration, we measured the discrimination performance (error rate = 1-accuracy) and computed the corresponding standard error of the mean across observers. Human performance in the vernier offset discrimination task was compared to the human data coming from the corresponding original crowding experiments (Figures 2 to 6).

Vernier offset matching algorithm

To avoid biases introduced by observers using different strategies to perform the mongrel discrimination tasks, we also performed mongrel vernier offset discrimination using a template matching algorithm. The algorithm searched for a target in the mongrels by sliding left- and right-sided vernier target templates over the whole image. For each location in the mongrel, a match value was defined by cross-correlating the template with the part of the image that lay under the template. Each match value was weighted by a function that decreased with the distance of the location of the template to the original position of the target, to help the algorithm focus on the most likely location of the vernier in the mongrel (Eq. 1).

$$M^s(i, j) = e^{-(D(i, j)/\sigma)^2} \cdot \sum_{k, l} T_{k, l}^s \cdot I_{i+k, j+l} \quad (1)$$

$M^s(i, j)$ was the weighted match value of the s -sided vernier template at location (i, j) , $T_{k, l}^s$ was the value of the s -sided vernier template at location (k, l) in the template coordinates, I was the mongrel array. $D(i, j)$ was the distance in pixels between the location of the template and the original target position and σ was the width of the weighting function in pixels. σ was set to 50 pixels. For each mongrel, the algorithm decided for a left or a right vernier as the side of the template that obtained the highest weighted match value.

Results

Lines experiment

In Manassi et al. (2012), crowding was strong when a vernier target was flanked on each side by two short lines or by two lines of the same length as the vernier, but weak when flanked by

two longer lines. When increasing the number of flankers, crowding decreased for short flankers, stayed constant with same-length flankers, and slightly decreased with long flankers (Figure 2, left). Hence, adding flankers can lead to non-monotonic effects in crowding strength, contrary to what is predicted by simple pooling models.

As with the simple pooling models, in both TTM validation tasks, crowding strength increased when increasing the number or the size of the flankers (Figure 2, center and right). The TTM performance differs from human data, in which adding flankers reduced crowding strength in certain conditions.

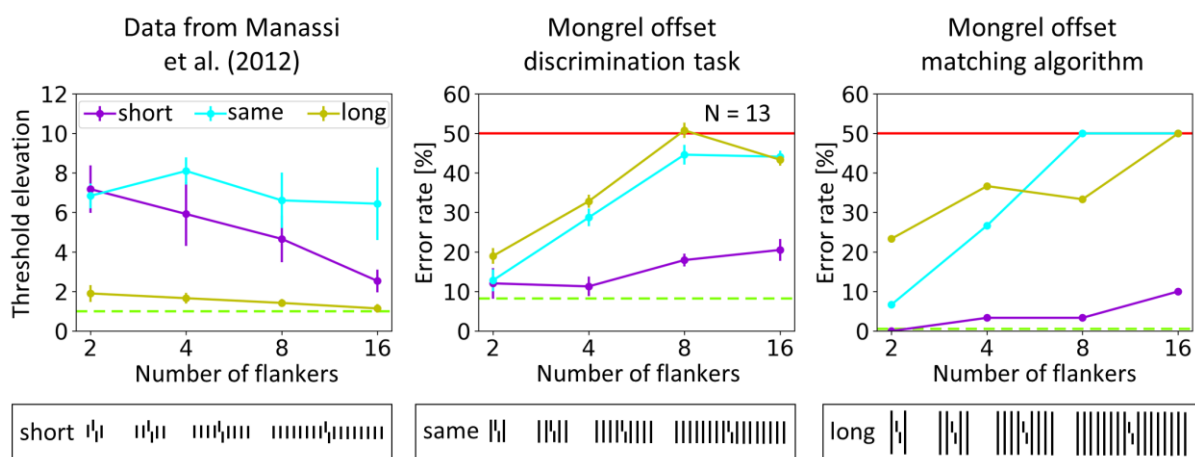


Figure 2. Lines. **Left.** Data from Manassi et al. (2012). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. **Center.** TTM validation in which observers discriminate between left and right offset verniers in mongrel images. **Right.** TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

Completion experiment

In Manassi et al. (2015), crowding was strong when a vernier was flanked by 16 same-length straight verniers but decreased when a same-length straight vernier mask was added at target location (Figure 3, left, straight vs comp16). Crowding was strong for control conditions in which a longer mask was used or using a same-length mask but having only 2 vernier flankers (Figure 3, left, comp16b & comp2). Hence, adding a single element can drastically change crowding strength, which cannot be explained by simple pooling models.

In both TTM validation tasks, crowding strength decreased when adding a same-length vernier mask at target location, as in the human data (Figure 3, center & right, straight vs comp16).

However, crowding strength also decreased when using a longer mask or having only 2 vernier flankers (Figure 3, center & right, straight vs comp16b & comp2), and gradually increased when adding more flankers ([Suppl. Inf. B, SB1](#)), showing that the configuration played no role.

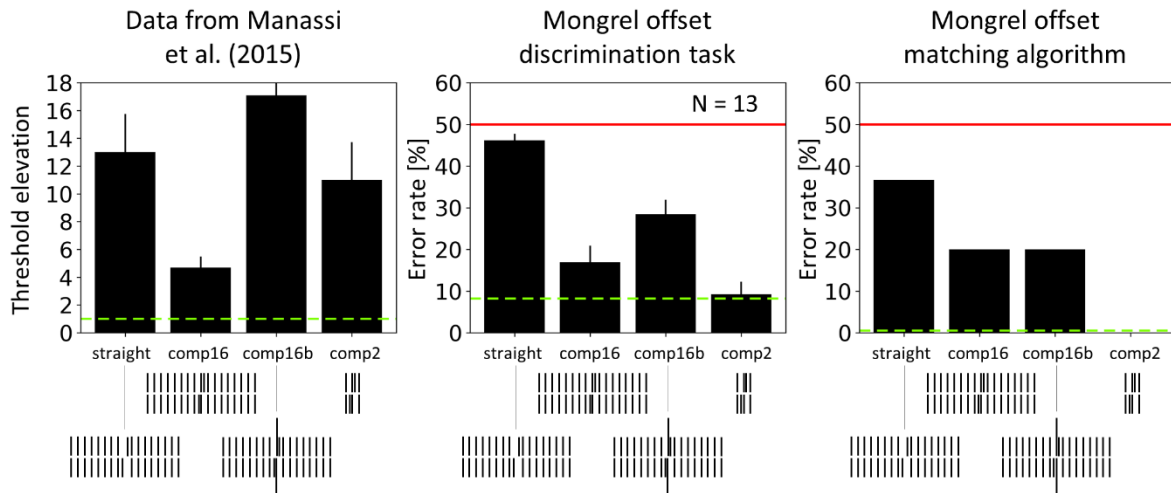


Figure 3. Completion. Left. Data from Manassi et al. (2015). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. **Center.** TTM validation in which observers discriminate between left and right offset verniers in mongrel images. **Right.** TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Note that the algorithm made 0% errors for in the comp2 condition (the data is not missing). Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

Boxes & crosses experiment

In Manassi et al. (2012), crowding was strong when the vernier target was flanked by 2 same-length flankers (Figure 4, left, bars). Crowding decreased when adding flankers to form boxes or boxes containing a cross (Figure 2, left, boxes and boxescrosses), but stayed high when the added flankers were not embedded in box shapes (Figure 2 left, crosses). These results were taken as evidence that flanker configuration modulates crowding strength.

The TTM failed to reproduce these results. In both TTM validation tasks, weak crowding was observed for the bars, and stronger crowding was observed when adding more flankers (Figure 2, center & right, bars vs boxes & crosses & boxescrosses), regardless of the configurations.

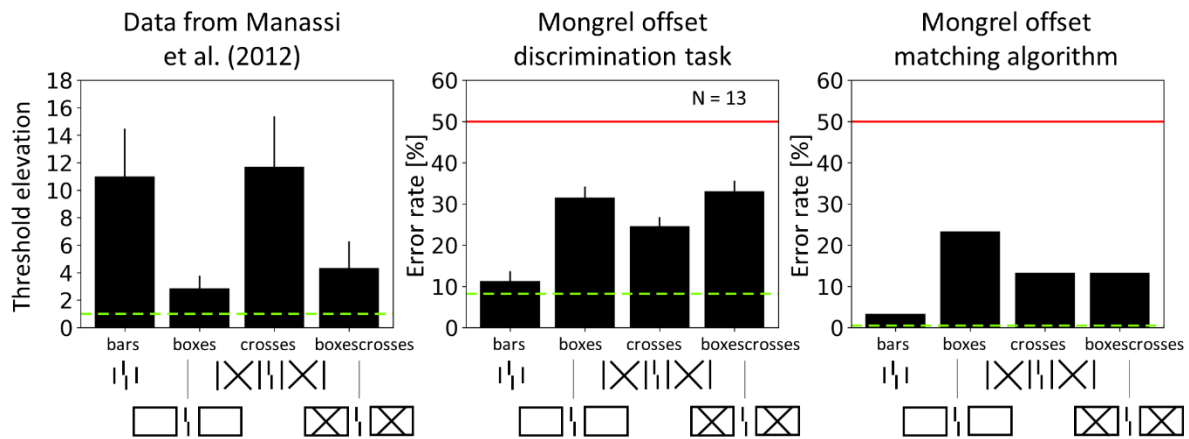


Figure 4. Boxes and crosses. **Left.** Data from Manassi et al. (2012). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. **Center.** TTM validation in which observers discriminate between left and right offset verniers in mongrel images. **Right.** TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note the different y-axis labels.

Shapes experiment

In Manassi et al. (2013), crowding was strong when the vernier target was flanked by a single square (Figure 5, left, 1S). Crowding decreased when the vernier was flanked by three additional squares on each side but remained strong when the added flankers were diamonds (Figure 5, left, 7S vs 7D1S). Crowding was strong in control conditions (Figure 5, left, 7L & 6L1S). The results showed that high-level shape processing can determine low-level vernier acuity.

The TTM did not reproduce this set of results. In both TTM validation tasks, crowding was strong for all tested conditions, independently of shape configuration (Figure 5, center & right). A similar pattern was found using diamonds instead of squares ([Suppl. Inf. B, SB2](#)).

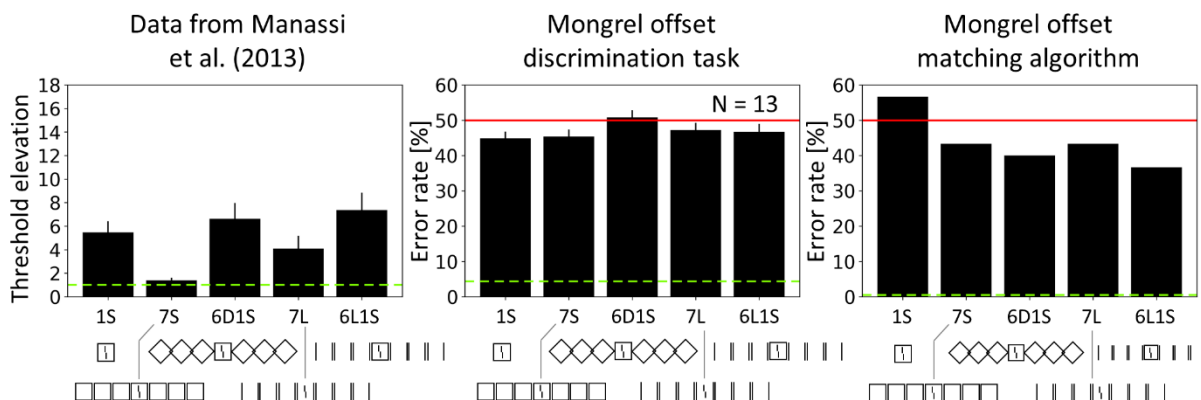


Figure 5. Shapes. **Left.** Data from Manassi et al. (2013). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 9 degrees of eccentricity. **Center.** TTM validation in which observers

discriminate between left and right offset verniers in mongrel images. **Right.** TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

Pattern experiment

In Manassi et al. (2016), crowding was strong when the vernier was embedded in a single square (Figure 6, left, 1S). Crowding was still strong when the vernier was embedded in an array of alternating squares and stars, but strongly decreased when the vernier was embedded in three identical rows of alternating squares and stars (Figure 6, left, A vs B). Crowding was strong in both control conditions (Figure 6, left, C & D). These results showed that the high-level spatial configurations of elements across large parts of the visual field, well beyond the range attributed to local pooling (Bouma, 1970), affect vernier discrimination performance.

Again, the TTM failed to reproduce these results. In both TTM validation tasks, crowding was strong for all tested conditions (Figure 6, center & right). Note that, to avoid ceiling effects in which crowding is too high to show differences between conditions, we also generated mongrels with a larger foveal radius (32 instead of 16 pixels) for all conditions in the Shapes and Patterns experiments (i.e., the ones in Figures 5 & 6, as well as [Suppl. Inf. B, SB2](#)). We also computed the TTM performance for these mongrels, using the template matching algorithm. We obtained lower crowding levels, but a similar qualitative behavior was observed ([Suppl. Inf. B, SB3](#)).

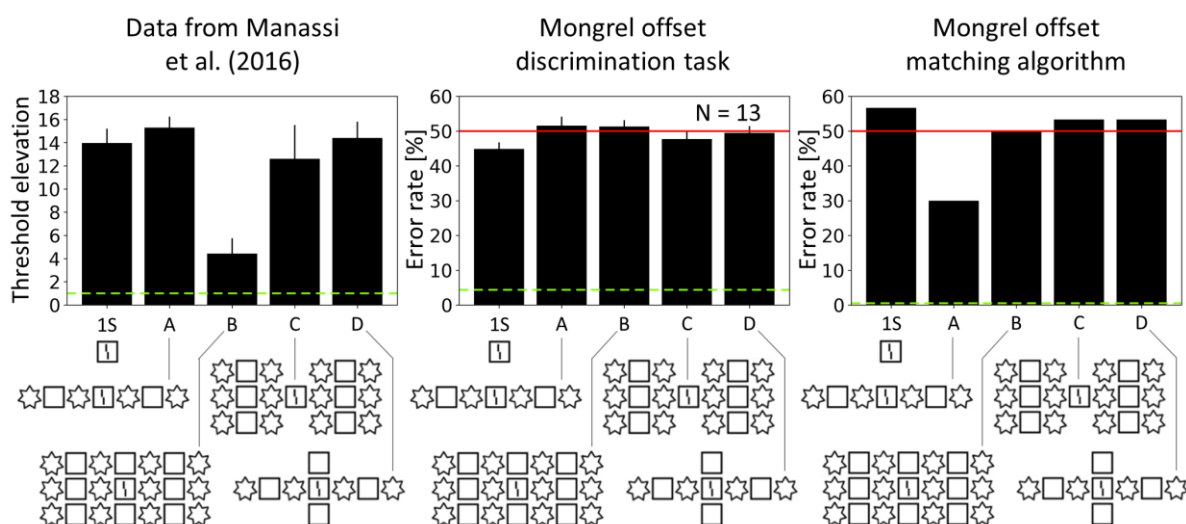


Figure 6: Patterns. **Left.** Data from Manassi et al. (2016). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 9 degrees of eccentricity. **Center.** TTM validation in which observers

discriminate between left and right offset verniers in mongrel images. **Right.** TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

Taken together, the results of the TTM matched none of the results of Manassi et al. (2012, 2013, 2015, 2016), which showed that: (1) increasing the number of flankers led to non-monotonic effects (Figure 2); (2) adding a single element drastically changed crowding behavior (Figure 3; completion effect); (3) flanker configuration determined crowding (Figure 4); (4) high-level processing determined low-level processing in crowding (Figure 5); (5) adding flankers beyond Bouma's window considerably affected crowding strength (Figure 6). None of these effects were reproduced by the TTM.

TTM & prediction power

As a global measure of the explanatory power of the TTM for each condition of Manassi et al. (2012, 2013, 2015, 2016), we plotted the error rates (%) in the mongrel vernier offset discrimination task as a function of the threshold elevation in the original crowding experiments (Figure 7A). The measured correlation was not significantly different from zero ($r(34) = -0.044$; $p\text{-value} = 0.799$), indicating that the TTM explains none of the reported results. A similar correlation was found using the template matching algorithm ([Suppl. Inf. B, SB5](#)).

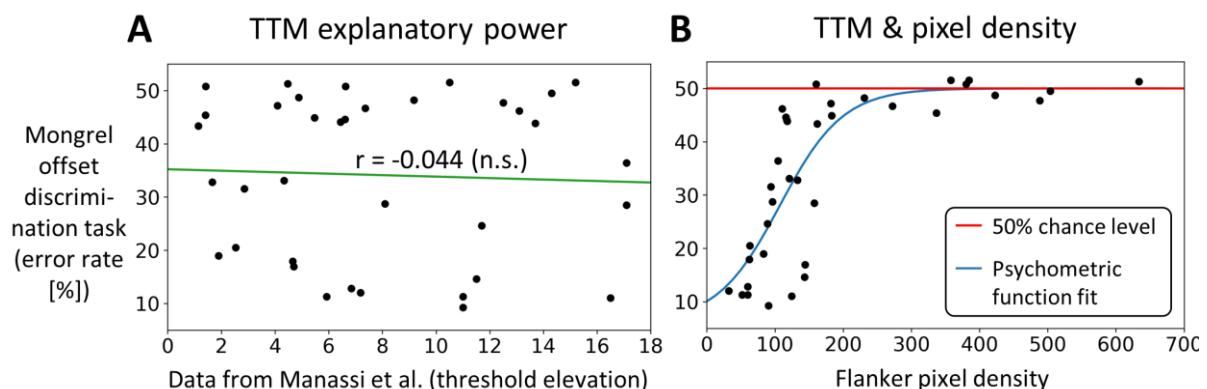


Figure 7. A. TTM performance in the mongrel vernier offset discrimination task showed no correlation ($r = -0.044$, $p = 0.799$, $BF_{01} = 4.672$) with the original data from Manassi et al. (2012, 2013, 2015, 2016). **B.** TTM performance as a function of the sum of the flanker pixels in the corresponding conditions. Each dot indicates a flanking condition in Figure 1. The red line indicates chance level performance. For illustrative reasons, we plotted all tested conditions in a unique graph. Separate plots for all experiments are shown in the supplementary information ([Suppl. Inf., SB6](#)). Fitting the data with a psychometric function (see Eq. 3 in [Suppl. Inf. B, SB11](#)), we found a strong correlation between the TTM and the fitted performance ($r(34) = 0.796$, $p < 0.001$, $BF_{10} > 10^6$).

Second, to assess the TTM behavior, we plotted its performance for each condition as a function of the flanker “density” in the corresponding original stimulus images (Figure 7B). To compute the flanker density, we counted the number of flanker pixels around the target. Each pixel contribution was weighted by a function that decreased with the distance to the target, mimicking Bouma’s law (Bouma, 1970). For each condition, the pixel density was defined as the sum of all weighted pixel contributions belonging to the flanker configuration (all details about the methods are given in [Suppl. Inf. B, SB11](#)). The error rate increased with flanker density (Figure 7B). Fitting the data with a psychometric function (see Eq. 3 in [Suppl. Inf. B, SB11](#)), we found a strong correlation between the TTM and the fitted performance ($r(34)=0.796$, $p<0.001$, $BF_{10}>10^6$). Crucially, this is the exact result that would be expected using a simple pooling model, suggesting that the TTM is blind to complex stimulus configuration and grouping cues, and simply relies on pixel density.

Grouping effects & target cueing

Rosenholtz et al. (2019) argued that the results in Manassi et al. (2012, 2013, 2015, 2016) do not necessarily imply the existence of grouping processes in crowding. Instead, it was proposed that target cueing plays a crucial role. Different stimulus configurations may cue the target location of the target in different ways, thus reducing target location uncertainty, leading to differences in crowding strength. Importantly, this explanation is entirely based on post-perceptual decision-making mechanisms. This is not a viable explanation for four main reasons.

First, cueing does not explain the results of Manassi et al. (2012, 2013, 2015, 2016). In these experiments, some flanker conditions strongly cue the target location but still produce strong crowding. In each comparison in Figure 8, the vernier target location is more cued by the flankers on the left side than on the right side. According to the cueing argument, crowding should be weaker on the left side compared to the right side. However, the human data show the exact opposite trend. For example, on the first line of the left panel in Figure 8, in the condition on the right (6S1D), the target location is clearly cued by the central diamond. There is no ambiguity at all about where the target is: it is inside the central diamond. In the condition on the left (7S), the line of squares casts more doubts on the location of the target. Nevertheless, crowding is 7.5 times larger on the right than on the left (Manassi et al., 2013).

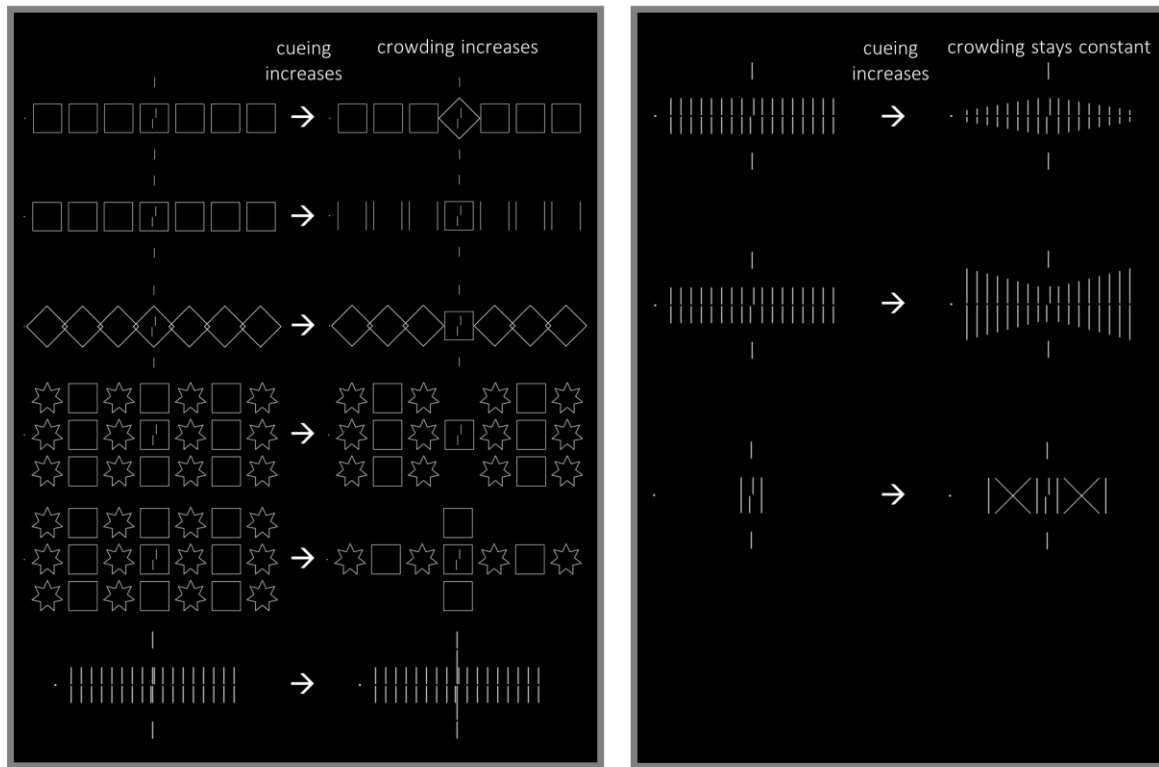


Figure 8. Right column, for both panels. Conditions in which the target location is weakly cued by the flanker configuration. Left Column, for both panels. Conditions in which the target location is strongly cued by the flanker configuration. If cueing had a strong impact on target discrimination performance, crowding would decrease from left to right in all comparisons. However, crowding strength either increases (left panel) or stays constant (right panel), while target cueing always increases. All conditions are taken from Manassi et al. (2012, 2013, 2015, 2016).

Second, in Manassi et al. (2012, 2013, 2015), two vertical lines were placed above and below the vernier target as “pointers”, in order to clearly cue the target location in all conditions. As reported in Manassi et al. (2012, 2013, 2015), the aim was to minimize the target location uncertainty. Rosenholtz et al. (2019) argued that these pointers may instead increase crowding by creating multiple offsets among vernier, flankers and pointers lines. However, in Manassi et al. (2012, 2013, 2015), the pointers were actually further from the vernier than reported by Rosenholtz et al. (2019), making this offset confusion argument unlikely (see Figure 17 in Rosenholtz et al. (2019) vs [Suppl. Inf. B, SB7](#)). Moreover, we measured the performance of the TTM model with all conditions, with or without pointers. The model did not show any significant increase in crowding strength with the pointers ([Suppl. Inf. B, SB8](#)).

Third, the effects measured in Manassi et al. (2012, 2013, 2015, 2016) correspond to changes in threshold elevation up to 10 times the unflanked threshold. The strength of cueing effects

in the literature has been consistently reported as small, with an average of 10% to 20% of difference in performance (Nazir, 1992; Scolari et al., 2007; Wilkinson et al., 1997; Yeshurun & Rashal, 2010). Thus, cueing does not seem even remotely sufficient to be considered as a viable explanation for global effects in crowding.

Fourth, a large part of these grouping effects in visual crowding were also found in foveal vision (Malania et al., 2007; Sayim et al., 2008, 2010; Waugh & Formankiewicz, 2020), where uncertainty is greatly reduced. Rosenholtz et al. (2019) argued that evidence for grouping effects in foveal vision casts doubts on whether these results are due to crowding. However, old and recent literature has shown evidence for crowding in foveal vision (Coates et al., 2013, 2018; Danilova & Bondarko, 2007; Flom et al., 1963; Lev et al., 2014; Lev & Polat, 2015; Sayim, Greenwood, et al., 2014; Siderov et al., 2013; Westheimer & Hauske, 1975), as well as grouping processes acting in foveal (Banks & White, 1984; Bock et al., 1993; Tannazzo et al., 2014) and peripheral vision (Banks & Prinzmetal, 1976; Banks & White, 1984; Livne & Sagi, 2007; Tannazzo et al., 2014; Wolford & Chambers, 1983). In other words, showing evidence for grouping effects in foveal vision does not invalidate any claim about grouping effects in crowding, but instead strengthens them.

To sum up, post-perceptual cueing cannot account for the effects measured in Manassi et al. (2012, 2013, 2015, 2016). These effects must hence be yielded by more complex interactions than what was previously thought to happen in visual crowding, such as contextual grouping (Malania et al., 2007; Manassi et al., 2012; Saarela et al., 2009).

TTM & Face Crowding

In the previous section, we showed that the TTM cannot explain the grouping effects found in Manassi et al. (2012, 2013, 2015, 2016) and that these effects cannot be explained by post-perceptual cueing. In this section, we tested the TTM with holistic face perception. Faces are considered as an invaluable tool to probe high-level visual processing, as they are analyzed holistically rather than as a set of separate features (Sergent, 1984). Mooney faces (Mooney, 1957), in particular, are the gold standard stimulus to test for holistic processing. Mooney faces (Fig. 9) are two-tone shadow images that are readily perceived as faces despite the lack of bottom-up processes that can segment or parse the image into features like an eye or mouth (Cavanagh, 1991; Fan et al., 2020; Grützner et al., 2010). That is, to see the mouth, eye, nose, eye separation, or other features, one must first recognize the stimulus as a face. This kind of holistic processing is necessary to recognize Mooney faces, and it has been operationalized in the literature by the inversion effect (McKone, 2004; Taubert et al., 2011): upright faces are recognized more easily than inverted ones (Farah et al., 1995; Kanwisher et al., 1998; Latinus & Taylor, 2005; Rossion, 2008; Sergent, 1984; Yin, 1969). The inversion effect is especially strong for Mooney faces (Canas-Bajo & Whitney, 2020; McKone, 2004; Schwiedrzik et al., 2018). Here, we tested the TTM with Mooney faces and found that it cannot predict two main results in holistic processing in crowding: (a) crowded object information is not lost at early stages of visual processing (inversion effect in a single face discrimination task; Bayle et al., 2011; Boucart et al., 2016; McKone, 2004) and (b) crowding occurs at high-level stages of visual processing between faces (crowding between holistic face representations; Farzin et al., 2009; Louie et al., 2007; Manassi & Whitney, 2018; Sun & Balas, 2015).

Methods

Single face discrimination task

We reproduced the single face discrimination task of Canas-Bajo & Whitney (2020). Observers were shown two images, one on each side of the visual field (Figure 9). Both images subtended a visual angle of 6° by 4.2° and were presented at the same eccentricity on both sides (6° , 10° , 14° or 18°). One image was always a face, whereas the other one was always a scrambled version of the same face (Schwiedrzik et al., 2018). The face could either be upright or inverted. Observers' task was to discriminate which of the two images was a face by pressing the left or

right arrow on a keyboard (2AFC), while fixating a cross in the center of the screen. The position on which the face appeared was randomized on each trial (either a face on the right and the corresponding scrambled face on the left or vice versa). There was no time constraint for giving a response. The distance to the screen was 64 cm.

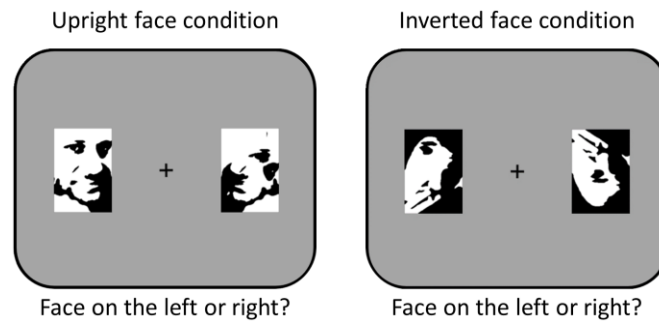


Figure 9. Single face discrimination task. Observers were asked to discriminate which of the two images was a face (left or right, 2AFC), by pressing the left or right arrow, while fixating the central cross. Across the experiment, the face could be either upright or inverted. In these examples, an upright face is presented on the left side (left panel), and an inverted face is presented on the right side (right panel).

There were 5 different faces, for a total of 20 different stimuli per eccentricity (2 sides, 2 face orientations, 5 different faces). Every stimulus was shown 10 times for a total of 200 trials per eccentricity. The experiment was run in blocks of fixed eccentricities. In each block, the stimuli were shown in a random left/right order. For each condition (upright vs inverted face) and eccentricity, we computed discrimination performance (error rate = 1-accuracy) and the corresponding standard error of the mean, computed over human observers (Figure 11A).

In order to validate the TTM, we tested mongrel images with the same single face discrimination task as in Canas-Bajo & Whitney (2020). For each stimulus, 10 different mongrels were generated using the TTM. Face discrimination performance in mongrel images was quantified by performing the single face discrimination task in free-viewing conditions. The experiment was run by blocks of eccentricity, for a total of 200 mongrels shown per eccentricity. Seven observers (2 males, 5 females, 25.4 ± 1.2 years old) performed the task. For each condition (upright vs inverted face) and eccentricity, we computed discrimination performance (accuracy [%]) and the corresponding standard error of the mean computed across observers. Performance in the single face discrimination task was then compared to the mongrel validation task (Figure 11).

Gender face discrimination task

Mongrel images were generated, following Experiment 6 from Farzin et al. (2009), which measured crowding induced by Mooney face flankers in a gender face discrimination task. Mooney faces were taken from Schwiedrzik et al. (2018). The size of the faces was the same as in Farzin et al. (2009), i.e., 1.53° by 2.48° . In these stimuli, the target face, which was always presented upright, could either be alone or surrounded by six other randomly selected Mooney faces (Figure 10). Flankers could either be upright or inverted. There were three different flanking conditions (target alone, upright flankers, inverted flankers) and four different target eccentricities (3° , 4.5° , 6° and 10°). Compared to the original experiment, we had an additional eccentricity (4.5°) in order to avoid floor and ceiling effects in the mongrel discrimination task. For each condition and eccentricity, 20 different Mooney faces were used as target (split equally between males and females), for a total of 240 original stimuli (20 faces x 3 flanking conditions x 4 eccentricities). 10 different mongrels were generated for each stimulus, for a total of 2400 unique samples shown to every participant. Seven observers (2 males, 5 females, 25.4 ± 1.2 years old) performed the task.

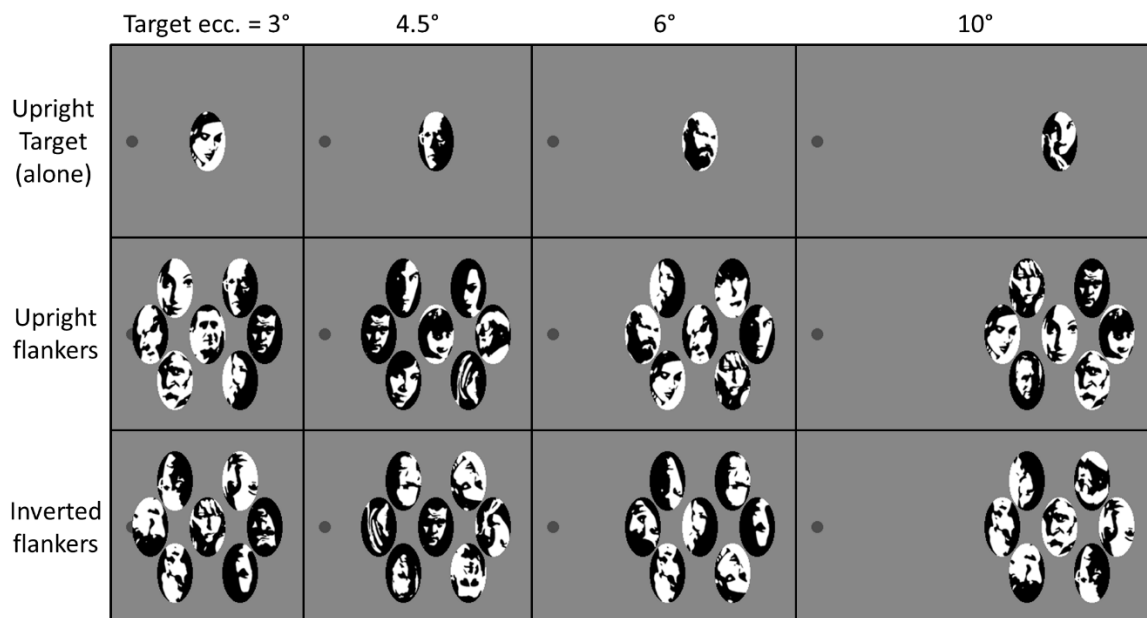


Figure 10. Examples of stimuli used in the face crowding task. There were three main conditions (upright target alone, target with upright flankers or target inverted flankers) presented at four different eccentricities.

Crowding strength in the TTM was quantified by performing a gender discrimination task in free-viewing conditions. We presented the generated mongrel images and asked observers to

indicate the gender of the target face (2AFC task). Mongrels were shown in a randomized order. Prior to the experiment, observers familiarized with the task as in the mongrel vernier offset discrimination task described above. For each condition and eccentricity, we computed the discrimination performance (accuracy [%]) and the corresponding standard error of the mean computed across observers. Performance in the mongrel gender crowding discrimination task was then compared to the behavioural data of Farzin et al. (2009; Figure 12).

In addition to the behavioral experiment, we measured the gender discrimination performance with a template matching algorithm. The algorithm matched original target face templates to all mongrel images. As for the Vernier offset matching algorithm, a face target was searched in the mongrels by sliding target face templates over the image (see Eq. 1 for the detailed computation). For each mongrel, the algorithm outputted the gender of the target face template that had the best match. Accuracy was computed as the percentage of correct answers. The performance of the algorithm was also compared to the data of Farzin et al. (2009; [Suppl. Inf. B, SB10](#)).

Results

Single face discrimination task

The results of the single face discrimination task are plotted in terms of accuracy (Figure 11A). Data were analyzed using a linear mixed effect model, with eccentricity and face orientation as the two fixed effects and individual subjects as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1)=0.062$, $p=0.803$). The main effect of face orientation was significant ($\chi^2(1)=30.99$, $p<0.001$), but not the effect of eccentricity ($\chi^2(1)=0.755$, $p=0.385$). The difference in effect size between the full model and the reduced model, excluding the effect of eccentricity, was 0.4% (full model: $r_m^2=0.243$, $r_c^2=0.696$, reduced model: $r_m^2=0.239$, $r_c^2=0.692$).

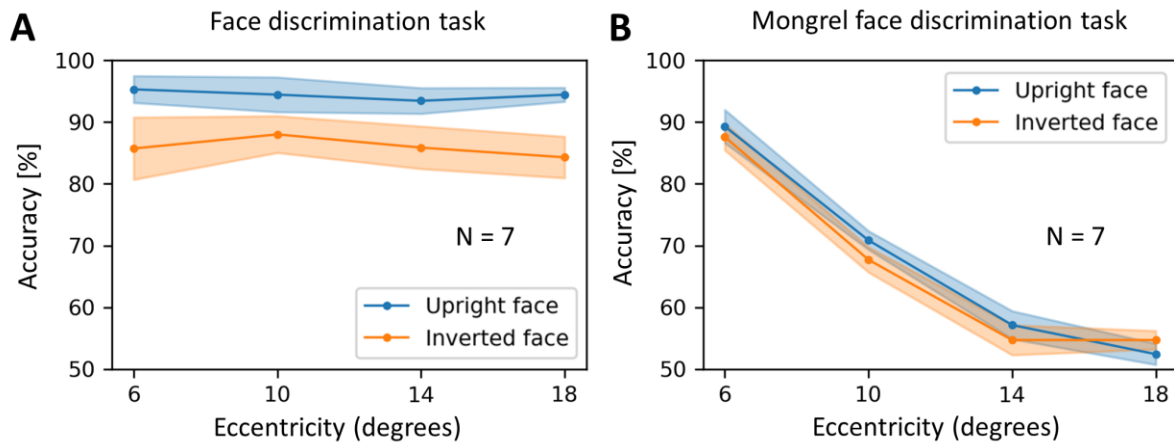


Figure 11. TTM & single Mooney face recognition. **A.** Face discrimination task. Observers were asked to discriminate an upright/inverted face from a scrambled face at all tested eccentricities. Accuracy remained on a constant high level for all eccentricities. Crucially, accuracy was higher for upright than for inverted faces. **B.** Mongrel face discrimination task. Accuracy decreased with increasing eccentricity, contrary to the behavioral results. Using a linear mixed effect model, no significant difference between the upright and inverted face conditions was observed (i.e., no significant effect of face orientation on model performance). Shaded regions indicate the standard error of the mean.

Observers were able to discriminate an upright/inverted face from a scrambled face at all tested eccentricities (Figure 11A). Crucially, observers' accuracy was higher for upright than inverted faces (Figure 11A, upright vs inverted), indicating a differential processing of inverted (low-level) and upright (holistic) faces, even at 18 deg eccentricity. The results suggest that face representations can survive any putative within-face low-level crowding, allowing holistic recognition of Mooney faces in the periphery.

Next, we tested whether the TTM could predict the inversion effect in individual Mooney faces (Figure 11B). As before, we validated the mongrels with the single face discrimination task. Observers were shown the mongrels of the original stimuli and were asked to tell which mongrel image was a face (free unconstrained viewing; see Methods for details). Data were analysed using a linear mixed effect model, with eccentricity and face orientation as the two fixed effects and individual subjects as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1)=0.647$, $p=0.421$). The main effect of eccentricity was significant ($\chi^2(1)=88.779$, $p<0.001$), but the effect of face orientation was not ($\chi^2(1)=0.494$, $p=0.482$). The difference in effect size between the full model, including both effects and the reduced model excluding the effect of face orientation, was only 0.2% (full model: $r_m^2=0.798$, $r_c^2=0.802$, reduced model: $r_m^2=0.796$, $r_c^2=0.800$).

These results show that the face discrimination performance in the TTM decreased with increasing eccentricity, contrary to the behavioral results (Figure 11, A vs B). More importantly, there was no difference between the upright and inverted mongrel face conditions (Figure 11B, orange vs. blue). The lack of inversion effect shows that the TTM treats upright and inverted faces as the same class of stimuli and, hence, it lacks any kind of holistic processing.

We ran another version of the mongrel validation task in which all mongrels generated with images comprising an inverted face were flipped upside-down. Hence, in this control task, observers were only shown upright mongrel faces, although they were processed either as upright or inverted faces in the TTM. This was done to isolate inversion effects in humans from inversion effects in the TTM as much as possible. The results were comparable ([Suppl. Inf. B, SB9](#)).

Taken together, the results show that holistic face recognition occurs also in peripheral vision, replicating and extending previous reports (Bayle et al., 2011; Boucart et al., 2016; Canas-Bajo & Whitney, 2020; McKone, 2004). Hence, crowded face-specific information is not lost at the early stages of visual processing but can be easily retrieved (Figure 11A). The TTM cannot explain this class of results. The TTM causes an irretrievable loss of face-specific information: discrimination performance drops with eccentricity and the inversion effect is eliminated (Figure 11B).

Gender face discrimination task

In Farzin et al. (2009), observers were asked to discriminate the gender of an upright face presented in the periphery. Accuracy decreased with increasing eccentricity (Figure 12A, black line). This decline in performance for isolated faces is an unsurprising consequence of the small size of the faces and the difficulty of the gender discrimination task. More importantly, when the same upright face was flanked by inverted or upright flankers, accuracy decreased, a standard hallmark of crowding. Crucially, upright flankers crowded more compared to inverted ones (blue line falls below orange line). This is an inversion effect in crowding: it shows that stimuli seen as faces crowd each other. When the same flanker stimuli are not seen as faces (i.e., are inverted), they do not crowd. Crowding is therefore gated by “similarity”, and the “similarity” must be at the level of holistic face representations. In the original publication (see Experiment 6 in Farzin et al., 2009), ANOVA resulted in a significant main effect of eccentricity

and flanker orientation (paired-samples 2-tailed t-tests revealed that upright face flankers impaired performance more than inverted flankers at 3° and 6° of eccentricity). Here we tested whether the TTM makes a similar prediction.

We computed the TTM performance for this experiment in a mongrel gender discrimination task (see Methods for details, gender face discrimination task). The results (Figure 12B) were analyzed using a linear mixed effect model, with eccentricity and face orientation (upright vs. inverted) as fixed effects and individual observers as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1)=0.479$, $p=0.489$). The main effect of eccentricity was significant ($\chi^2(1)=121.11$, $p<0.001$), but the effect of face orientation was not ($\chi^2(1)=0.620$, $p=0.431$). The difference in effect size between the full model, including both effects (eccentricity and face orientation) and the reduced model excluding the effect of face orientation, was only 0.2% (full model: $r_m^2=0.691$, $r_c^2=0.691$, reduced model: $r_m^2=0.689$, $r_c^2=0.689$).

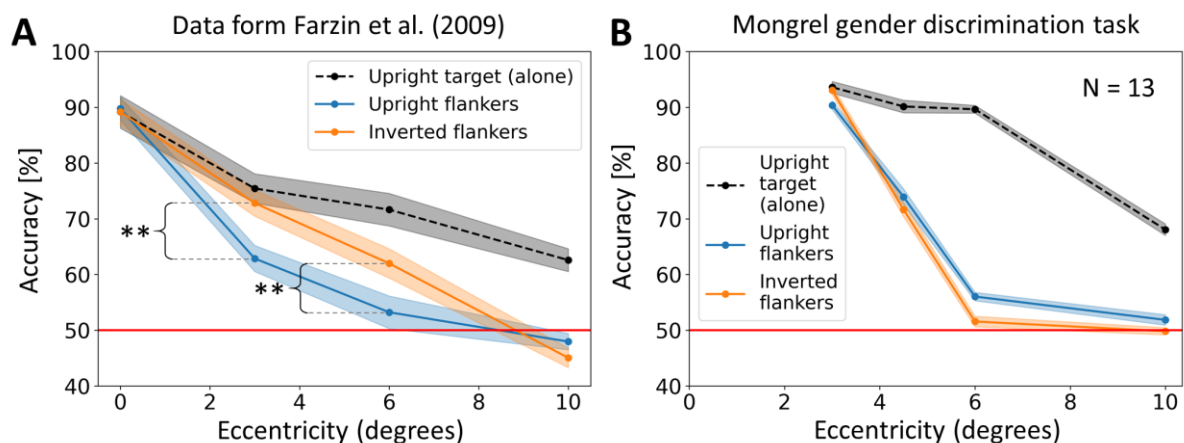


Figure 12. TTM & crowding of Mooney faces. **A.** Face crowding task, data from Farzin et al. (2009). Target discrimination performance decreased when eccentricity increased. When the target face was flanked by inverted faces, crowding increased with increasing eccentricity (orange). When the target was flanked by upright faces, crowding increased even more with eccentricity (blue). Shaded regions indicate the standard error of the mean. Stars indicate a significant difference in crowding strength between the upright and inverted flanker face conditions (paired student t-test, 2-tails). **B.** Mongrel face crowding task. Accuracy decreased with eccentricity. When analysing the results using a linear mixed effect model, no effect of flanker face orientation was exposed. Shaded regions indicate the standard error of the mean.

As in Farzin et al. (2009; Figure 12A), TTM performance decreased with eccentricity (Figure 12B). However, unlike Farzin, et al (2009), the linear mixed effect model revealed no significant

overall effect of flanker orientation, and no interaction between eccentricity and target orientation. Simply put, the TTM does not predict a systematic difference in crowding as a function of the flanker orientation. And, when TTM does predict a trending difference, it is often in a direction opposite that in the empirical data (blue-above-orange in Fig. 12B compared to orange-above-blue in Fig. 12A). These results show that the TTM can predict a general increase of crowding with eccentricity (i.e., low-level crowding) but it fails to predict face-selective or holistic effects in crowding.

Taken together, the results depicted in Figure 11 and 12 show that the TTM is not able to predict peripheral face recognition or the effects of high-level face processing in crowding. It fails to predict crowding of single faces (Figure 11) and multiple faces (Figure 12). In fact, target information in the TTM is irretrievably lost at a low-level pooling stage and crowding occurs only between low-level features (Figure 7). In this light, it is unsurprising that the TTM fails to explain a broad array of findings in the peripheral face recognition literature (Boucart et al., 2016; Farzin et al., 2009; Kovács et al., 2017; Kreichman et al., 2020).

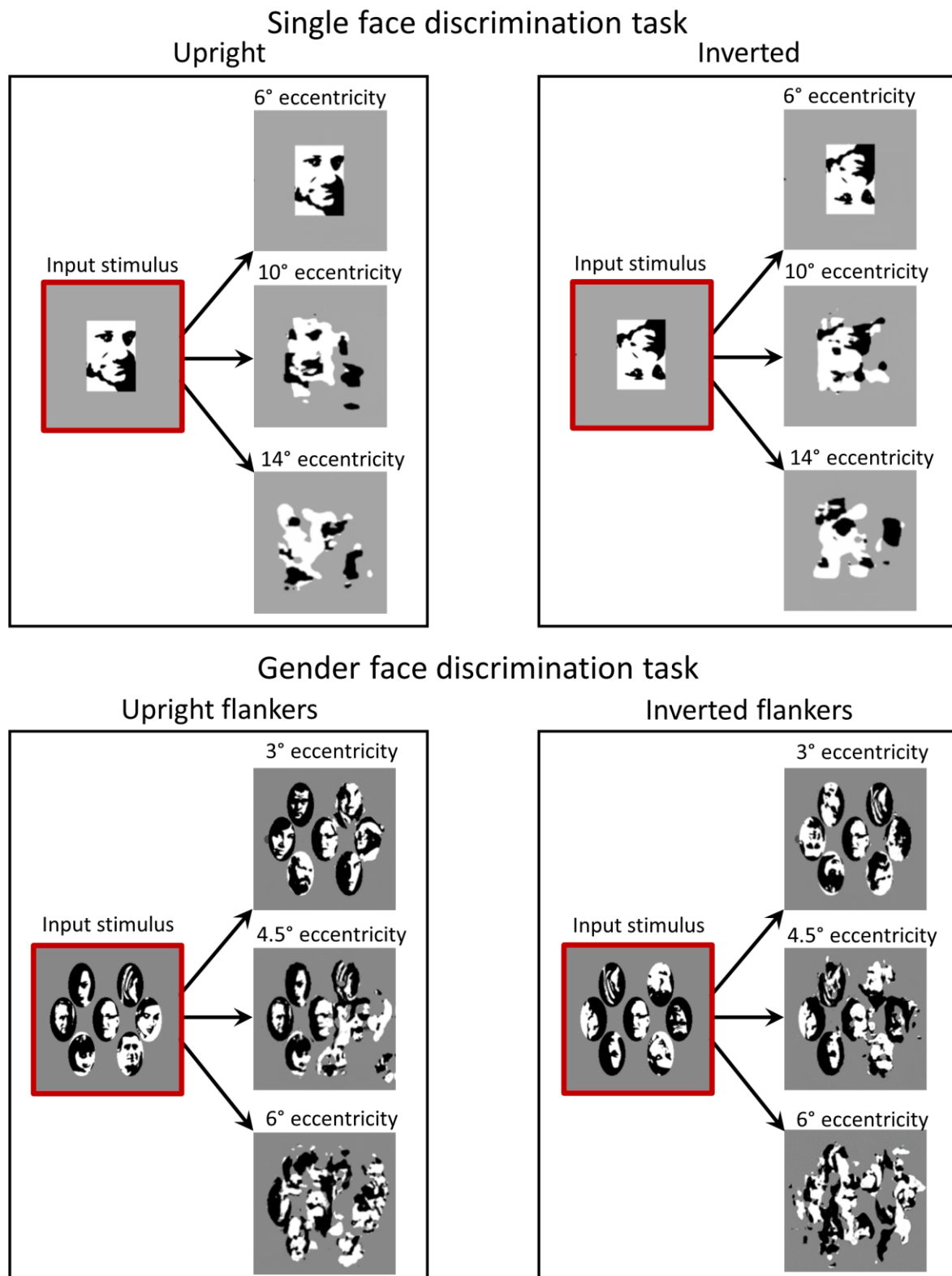


Figure 13. TTM mongrel examples used in the single face and gender face discrimination tasks. The stimuli (TTM input) are highlighted in red. To give a representative sample of the TTM outputs for each example, we show mongrels for different eccentricities. Note that we cropped the mongrels for ease of comparison. All mongrels can be found at https://github.com/albornet/TTM_Verniers_Faces_Mongrels.

Discussion

Classic models describe crowding as a relatively simple, local and low-level phenomenon (Greenwood et al., 2009; Levi et al., 2002; Nandy & Tjan, 2012; Parkes et al., 2001; Van den Berg et al., 2010; Wilkinson et al., 1997). Recent studies, however, provided clear-cut psychophysical evidence that crowding is in fact more complex than previously thought, involving global interactions and occurring at multiple stages of visual processing (Farzin et al., 2009; Manassi et al., 2012, 2013, 2015, 2016; Manassi & Whitney, 2018; Saarela et al., 2009, 2010; Whitney & Levi, 2011). Against this new view of crowding, Rosenholtz et al. (2019) argued that (1) high-dimensional pooling is sufficient to explain the new results and (2) target cueing plays a crucial role in these effects. Here, we quantitatively tested these claims on a large array of experimental data and showed that (1) TTM fails to account for human crowding performance and (2) target cueing does not play a role. In the following, we will describe implications from our two sets of data on grouping effects and face recognition.

TTM & grouping effects

Using a mongrel offset discrimination task, we showed that the TTM did not reproduce any of the results of Manassi et al. (2012, 2013, 2015, 2016), in which: (1) increasing the number of flankers sometimes reduces crowding strength (Figure 2); (2) adding a single element has a dramatic effect on crowding strength (Figure 3; completion effect); (3) the overall configuration of the flankers determines crowding (Figure 4); (4) high-level processing strongly affects low-level processing (Figure 5), and (5) adding flankers beyond Bouma's window strongly modulates crowding strength (Figure 6).

It was proposed that the best predictor of visual crowding is grouping between target and flankers: crowding increases when the target groups with the flankers, but decreases when the target ungroups and stands out from the flankers (Malania et al., 2007; Saarela et al., 2009, 2010; Sayim et al., 2008, 2010). In line with this hypothesis, in Manassi et al. (2012) and in Saarela et al. (2009), subjective ratings on target-flankers grouping correlated with crowding strength. Furthermore, Doerig et al. (2019) showed that only models that included a grouping stage could explain these results (see also Doerig, Schmittwilken, et al., 2020). In the TTM, crowding strength was never reduced, when additional flankers were added, regardless of flanker configuration (Figures 2-6).

The only result that was reproduced by the TTM is the reduction in crowding strength when adding a straight-vernier mask at target location in the Completion experiment (Figure 3, center and right, straight vs comp16). We attribute this reduction in crowding strength to a *local* effect of the mask. When the mask is added, the region around the target is summarized by different local statistics than when the mask is absent (higher spatial frequencies, locally). Hence, this region stands out from the rest of the image. It is thus better reconstructed by the TTM, yielding better performance. However, crowding in the TTM was still reduced in the control conditions (Figure 3, center and right, comp16b & comp2), further supporting the notion that the mask induces a local effect only: when the configuration of the grating is broken by the presence of the long mask (comp16b) or by the absence of many flankers (comp2), crowding is still reduced. This is in contradiction to the human data, in which crowding is reduced by the global layout of the flankers. In addition, crowding strength with various numbers of same length flankers ([Suppl. Inf. B, SB1](#)), was always weaker with than without the mask and always increased with more flankers, contrary to the human data.

Taken together, these results suggest that a pooling model, even a high-dimensional one, cannot account for the complexity of visual crowding. Comparing the performance of the TTM for all tested conditions to the corresponding human performance measured in Manassi et al (2012, 2013, 2015, 2016), we found no significant correlation (Figure 7A). Moreover, we found that the TTM performance strongly correlates with the amount of flankers around the target (Figure 7B), similar to a simple pooling model. It seems as if the TTM is blind to complex configurations and grouping cues. We propose that the reason for this lies in the model architecture, i.e., feedforward pooling cannot explain high-level effects in crowding (Doerig, Bornet, et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken, et al., 2020).

There are several reasons why the TTM failed. First, elements outside the pooling regions of the TTM can change crowding performance in humans but not in the TTM. Second, the strength of the TTM is the compression of information implemented by the computation of summary statistics, which may play a role for grouping. However, the TTM does not allow to change the scale of the pooling regions in function of the specificities of the stimuli. For this reason, the TTM filters out fine-grained information that is crucial for human performance. As put by Wallis et al. (2017), “Based on our experiments we speculate that the concept of summary statistics cannot fully account for peripheral scene appearance. Pooling in fixed

regions will either discard (long-range) structure that should be preserved or preserve (local) structure that could be discarded. Rather, we believe that the size of pooling regions needs to depend on image content". We think that the TTM summary statistics are important in crowding but need to adapt to the stimulus global configuration (including feedback processing) and not hard-wired.

Importantly, in contrast to what was proposed by Rosenholtz et al. (2019), cueing cannot account for grouping effects in crowding. Cueing may be an explanation for some configurations, but overall, it is a poor predictor of crowding strength (Figure 8). Moreover, cueing studies only report small effect sizes (Nazir, 1992; Scolari et al., 2007; Yeshurun & Rashal, 2010), far beneath the effect sizes measured in Manassi et al. (2012, 2013, 2015, 2016). Hence, grouping effects in crowding are not post-perceptual, e.g., caused by differences in target visibility or target cueing. They are purely perceptual and are caused by complex target-flanker interactions occurring along the visual processing hierarchy.

Rosenholtz et al. (2019) argued that, since effects of contextual grouping were also found in foveal vision (Saarela & Herzog, 2008; Sayim et al., 2010, 2011; Sayim, Manassi, et al., 2014; Waugh & Formankiewicz, 2020), they may not be due to genuine crowding. However, literature showed that crowding can occur in foveal (Coates et al., 2013, 2018; Danilova & Bondarko, 2007; Flom et al., 1963; Lev et al., 2014; Lev & Polat, 2015; Sayim, Greenwood, et al., 2014; Siderov et al., 2013; Westheimer & Hauske, 1975) and peripheral vision (Levi, 2008; Pelli, 2008). Importantly, the stimuli in foveal experiments were the same as in peripheral crowding and so were the results. In any case, the TTM needs either to explain the peripheral effects, independent of where or not there is foveal crowding, or to convincingly explain why not.

TTM & face crowding

In another set of experiments (Figures 9 and 11), we focused on single face recognition in peripheral vision. Using a single Mooney face discrimination task, we showed that holistic face recognition occurs in peripheral vision, i.e., a better recognition performance for upright than for inverted faces (Figure 11A, upright vs inverted), reproducing the results found in Canas-Bajo & Whitney (2020) and in line with old and recent literature (Farah et al., 1995; Rossion, 2008; Sergent, 1984; Yin, 1969). The advantage in recognizing upright Mooney faces speaks

for a differential processing involved between inverted (low-level) and upright (holistic) faces. These results cannot be explained by models of crowding based on simple pooling. According to this class of models, the two-tone black and white blobs constituting a Mooney face should crowd themselves in peripheral vision (e.g., Fig. 11B), thus becoming more unrecognizable when increasing in eccentricity (Martelli et al., 2005). Instead, our results show that the representation of these object parts nevertheless survives crowding (see also Manassi & Whitney, 2018), allowing holistic recognition of Mooney faces.

Using a mongrel Mooney face discrimination task, we showed that the low-level visual information that is merged in the pooling stage of the TTM is irretrievably lost. Despite the high dimensionality of the pooling in the TTM, at increasing eccentricities the features that compose the faces crowd each other in the model and cannot be used for further processing in the mongrel face discrimination task (Figure 11B). This is in contradiction with the results of the single face discrimination task we performed (Figure 11A; Canas-Bajo & Whitney (2020), and with recent evidence that face representations can survive crowding and influence subsequent perceptual judgments (Kouider et al., 2011).

Next, we focused on holistic face crowding (as found in Experiment 6 of Farzin et al., 2009; Figure 12A), in which upright flanker faces yielded more crowding than inverted ones in a gender face discrimination task. This inversion effect showed that crowding can occur selectively between high-level holistic representations conveyed by Mooney faces.

We tested whether the TTM could predict this result. Using a mongrel gender crowding discrimination task (Figure 10), we showed that the TTM did not reproduce holistic face crowding (Figure 12B). While crowding occurred in the TTM when face flankers were added, there was no effect of flanker face orientation on the TTM performance. This result confirms that crowding indeed happens selectively between high-level representations and cannot arise from low-level accounts, even using a high-dimensional pooling stage.

It was recently argued that the face crowding results in Farzin et al. (2009) may be due to differences in flankers reportability (Reuther & Chakravarthi, 2019). When target and flankers belong to the same category (upright faces as target and flankers), crowding may arise in part from reporting the flankers' gender instead of the target one (substitution errors). However, when target and flankers belong to different categories (upright face as target and inverted

faces as flankers), substitution errors are less likely to occur because flankers cannot be inadvertently reported. Hence, the decrease in crowding strength may be ascribed to the lack of substitution errors. As in the target cueing argument (Figure 8), this explanation assumes that target location uncertainty (and substitution errors, as a consequence) plays a crucial role in crowding, driving the entire difference in crowding strength between upright and inverted face flankers. We would argue that the stimuli in Farzin et al. (2009) are edge-defined high-contrast faces with clearly defined target locations, and thus make it unlikely that so many substitution errors occur on an object level. More importantly, however, this argument assumes that, prior to target-flanker substitution, upright/inverted faces are processed differently, thus implying some kind of holistic face processing, just as Farzin et al. (2009) suggested.

Model assessment method

It may be argued that the TTM may reproduce high-level effects in crowding using a different set of model parameters. For example, some of the TTM failures could result from ceiling effects. Here, using the parameters suggested by Rosenholtz et al. (2019), we found that crowding was too weak for most stimuli, which may obscure complex effects. For this reason, we decreased the fovea radius parameter from 32 to 16 pixels to increase crowding in all conditions. Still, for most stimuli that included large flanker configurations at large eccentricities (Shapes and Patterns experiments; Figures 5 & 6, as well as [Suppl. Inf. B, SB2](#)), performance was at chance level and hence, high-level effects might have gone unnoticed. For all these stimuli, we ran a follow-up experiment in which we kept the fovea radius parameter as 32 pixels to make the task easier. This did not improve the model predictions, as measured by the template matching algorithm ([Suppl. Inf. B, SB3](#)).

Moreover, it may be argued that assessing the TTM performance using behavioral mongrel discrimination tasks can introduce biases coming, for example, from different strategies used by human observers. First, it should be noted that the method we used is the same as in Rosenholtz et al. (2019) and their previous work (Balas et al., 2009; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, 2011; Zhang et al., 2015). Nevertheless, to control for unwanted human biases, we also quantified performance using a template matching algorithm (see Methods for details). This did not change the results qualitatively

(Figures 2 to 6, as well as [Suppl. Inf. B, SB1-5 and SB10.](#)). The measured performances were similar to what was measured in the behavioral tasks, and none of the high-level effect of crowding were reproduced.

Model improvements

We would like to mention that the TTM accounts for a variety of perceptual properties of human vision (Alexander et al., 2014; Chang & Rosenholtz, 2016; Rosenholtz, 2011; Rosenholtz, Huang, Raj, et al., 2012), as well as many properties of crowding (substitution effects, Bouma's window, etc.). Hence, our results should not be taken as a complete invalidation of the model. They rather suggest that, to capture human behavior fully, models of crowding and of vision in general need to incorporate more specific mechanisms that account for complex visual processing. Our results provide evidence that high-level effects cannot emerge even from the most sophisticated and high-dimensional pooling models, such as the TTM.

How could these models be improved? First, to explain the complex effects in Manassi et al (2012, 2013, 2015, 2016), we propose to add a recurrent grouping and segmentation stage to existing models of crowding. In such models, the high-level configuration of the stimulus affects lower-level target acuity, so that crowding interference only occurs within perceptual groups. Recent work confirmed that recurrent grouping and segmentation processes are a promising addition to capture global aspects of crowding (Bornet et al., 2019; Doerig, Bornet, et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken, et al., 2020; Francis et al., 2017). Second, to explain why crowding happens at multiple levels, such as in holistic crowding between faces (Farzin et al., 2009; Manassi & Whitney, 2018; Whitney & Levi, 2011), we propose to consider high-level statistics in high-dimensional pooling models, such as the TTM. By pooling information at all stages (instead of a low-level unique one), the model could account for holistic effects in high-level crowding. Alternatively, Chaney et al. (2014) proposed the Hierarchical Sparse Selection (HSS) model. In this model, fine-grained information is preserved by the feature integration process occurring in the visual cortex because of the high density of neurons paving the visual field (note that this is slightly different to the high-dimensional pooling stage of the TTM, in which fine-grained information is preserved because of the large number of pooled features). Crowding happens in the HSS model because, for the

sake of efficient visual perception, the neurons that are selected to decode the target features are sampled sparsely.

In conclusion, our results provide evidence that high-level effects cannot emerge even from the most sophisticated and high-dimensional pooling models, such as the TTM. Moreover, target cueing is not a viable explanation for these effects. Hence, crowding remains a complex, global and multi-level perceptual phenomenon, as well as a precious and versatile probe to understand what may be missing from current models of human vision.

References

- Alexander, R. G., Schmidt, J., & Zelinsky, G. J. (2014). Are summary statistics enough? Evidence for the importance of shape in guiding visual search. *Visual Cognition*, 22(3–4), 595–609.
- Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research*, 16(1), 71–78. [https://doi.org/10.1016/0042-6989\(76\)90078-X](https://doi.org/10.1016/0042-6989(76)90078-X)
- Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, 9(12), 13–13.
- Banks, W. P., & Prinzmetal, W. (1976). Configurational effects in visual information processing. *Perception & Psychophysics*, 19(4), 361–367. <https://doi.org/10.3758/BF03204244>
- Banks, W. P., & White, H. (1984). Lateral interference and perceptual grouping in visual detection. *Perception & Psychophysics*, 36(3), 285–295.
- Bayle, D. J., Schoendorff, B., Hénaff, M.-A., & Krolak-Salmon, P. (2011). Emotional facial expression detection in the peripheral visual field. *PloS One*, 6(6), e21584.
- Bock, J. M., Monk, A. F., & Hulme, C. (1993). Perceptual grouping in visual word recognition. *Memory & Cognition*, 21(1), 81–88. <https://doi.org/10.3758/BF03211167>
- Bornet, A., Kaiser, J., Kroner, A., Falotico, E., Ambrosano, A., Cantero, K., Herzog, M. H., & Francis, G. (2019). Running large-scale simulations on the NeuroRobotics Platform to understand vision-the case of visual crowding. *Frontiers in NeuroRobotics*, 13, 33.
- Boucart, M., Lenoble, Q., Quettelart, J., Szaffarczyk, S., Despretz, P., & Thorpe, S. J. (2016). Finding faces, animals, and vehicles in far peripheral vision. *Journal of Vision*, 16(2), 10–10.
- Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, 226(5241), 177–178.
- Canas-Bajo, T., & Whitney, D. (2020). Stimulus-specific individual differences in holistic perception of Mooney faces. *Frontiers in Psychology*, 11.
- Cavanagh, P. (1991). What's up in top-down processing. *Representations of Vision: Trends and Tacit Assumptions in Vision Research*, 295–304.
- Chaney, W., Fischer, J., & Whitney, D. (2014). The hierarchical sparse selection model of visual crowding. *Frontiers in Integrative Neuroscience*, 8.
- Chang, H., & Rosenholtz, R. (2016). Search performance is better predicted by tileability than presence of a unique basic feature. *Journal of Vision*, 16(10), 13–13.
- Chung, S. T., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research*, 41(14), 1833–1850.
- Coates, D. R., Chin, J. M., & Chung, S. T. (2013). Factors affecting crowded acuity: Eccentricity and contrast. *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, 90(7).
- Coates, D. R., Levi, D. M., Touch, P., & Sabesan, R. (2018). Foveal crowding resolved. *Scientific Reports*, 8(1), 1–12.
- Danilova, M. V., & Bondarko, V. M. (2007). Foveal contour interactions and crowding effects at the resolution limit of the visual system. *Journal of Vision*, 7(2), 25–25.
- Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. Global processing in humans and machines. *Vision Research*, 167, 39–45.
- Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., & Herzog, M. H. (2019). Beyond Bouma's window: How to explain global aspects of crowding? *PLoS Computational Biology*, 15(5), e1006580.
- Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLOS Computational Biology*, 16(7), e1008017.
- Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision*, 16(2), 13–13.
- Fan, X., Wang, F., Shao, H., Zhang, P., & He, S. (2020). The bottom-up and top-down processing of faces in the human occipitotemporal cortex. *ELife*, 9, e48764.
- Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, 21(3), 628.

- Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision*, 9(6), 18–18.
- Flom, M. C., Heath, G. G., & Takahashi, E. (1963). Contour interaction and visual resolution: Contralateral effects. *Science*, 142(3594), 979–980.
- Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review*, 124(4), 483.
- Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences*, 106(31), 13130–13135.
- Grützner, C., Uhlhaas, P. J., Genc, E., Kohler, A., Singer, W., & Wibral, M. (2010). Neuroelectromagnetic correlates of perceptual closure processes. *Journal of Neuroscience*, 30(24), 8342–8352.
- Harrison, W. J., Retell, J. D., Remington, R. W., & Mattingley, J. B. (2013). Visual crowding at a distance during predictive remapping. *Current Biology*, 23(9), 793–798.
- Herzog, M. H., & Manassi, M. (2015). Uncorking the bottleneck of crowding: A fresh look at object recognition. *Current Opinion in Behavioral Sciences*, 1, 86–93. <https://doi.org/10.1016/j.cobeha.2014.10.006>
- Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision*, 15(6), 5–5.
- Herzog, M. H., Sayim, B., Manassi, M., & Chicherov, V. (2016). What crowds in crowding? *Journal of Vision*, 16(11), 25–25.
- Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, 68(1), B1–B11.
- Keshvari, S., & Rosenholtz, R. (2016). Pooling of continuous features provides a unifying account of crowding. *Journal of Vision*, 16(3), 39–39. <https://doi.org/10.1167/16.3.39>
- Kimchi, R., & Pirkner, Y. (2015). Multiple level crowding: Crowding at the object parts level and at the object configural level. *Perception*, 44(11), 1275–1292.
- Kouider, S., Berthet, V., & Faivre, N. (2011). Preference is biased by crowded facial expressions. *Psychological Science*, 22(2), 184–189.
- Kovács, P., Knakker, B., Hermann, P., Kovács, G., & Vidnyánszky, Z. (2017). Face inversion reveals holistic processing of peripheral faces. *Cortex*, 97, 81–95.
- Kreichman, O., Bonne, Y. S., & Gilaie-Dotan, S. (2020). Investigating face and house discrimination at foveal to parafoveal locations reveals category-specific characteristics. *Scientific Reports*, 10(1), 1–15.
- Latinus, M., & Taylor, M. J. (2005). Holistic processing of faces: Learning effects with Mooney faces. *Journal of Cognitive Neuroscience*, 17(8), 1316–1327.
- Lev, M., & Polat, U. (2015). Space and time in masking and crowding. *Journal of Vision*, 15(13), 10–10.
- Lev, M., Yehezkel, O., & Polat, U. (2014). Uncovering foveal crowding? *Scientific Reports*, 4, 4067.
- Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, 48(5), 635–654.
- Levi, D. M., Hariharan, S., & Klein, S. A. (2002). Suppressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of Vision*, 2(2), 3–3.
- Levi, D. M., Toet, A., Tripathy, S. P., & Kooi, F. L. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision*, 8(2), 255–279.
- Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision*, 7(2), 4–4.
- Livne, T., & Sagi, D. (2010). How do flankers’ relations affect crowding? *Journal of Vision*, 10(3), 1–1.
- Louie, E. G., Bressler, D. W., & Whitney, D. (2007). Holistic crowding: Selective interference between configural representations of faces in crowded scenes. *Journal of Vision*, 7(2), 24–24.
- Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- Malania, M., Herzog, M. H., & Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *Journal of Vision*, 7(2), 1–1.
- Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision*, 15(8), 16–16.

- Manassi, M., Lonchampt, S., Clarke, A., & Herzog, M. H. (2016). What crowding can tell us about object representations. *Journal of Vision*, 16(3), 35–35.
- Manassi, M., Sayim, B., & Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision*, 12(10), 13–13.
- Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision*, 13(13), 10–10.
- Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology*, 28(3), R127–R133.
- Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision*, 5(1), 6–6.
- McKone, E. (2004). Isolating the special component of face recognition: Peripheral identification and a Mooney face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 181.
- Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 11(4), 219.
- Nandy, A. S., & Tjan, B. S. (2012). Saccade-confounded image statistics explain visual crowding. *Nature Neuroscience*, 15(3), 463–469. <https://doi.org/10.1038/nn.3021>
- Nazir, T. A. (1992). Effects of lateral masking and spatial precueing on gap-resolution in central and peripheral vision. *Vision Research*, 32(4), 771–777.
- Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4(7), 739–744.
- Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology*, 18(4), 445–451. <https://doi.org/10.1016/j.conb.2008.09.008>
- Reuther, J., & Chakravarthi, R. (2019). Response selection modulates crowding: A cautionary tale for invoking top-down explanations. *Attention, Perception, & Psychophysics*, 1–16.
- Rosenholtz, R. (2014). Texture perception. *Oxford Handbook of Perceptual Organization*, 167, 186.
- Rosenholtz, R. (2011). What your visual system sees where you are not looking. *Human Vision and Electronic Imaging XVI*, 7865, 786510.
- Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology*, 3, 13.
- Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, 12(4), 14–14. <https://doi.org/10.1167/12.4.14>
- Rosenholtz, R., Yu, D., & Keshvari, S. (2019). Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of Vision*, 19(7), 15–15.
- Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica*, 128(2), 274–289.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237.
- Saarela, T. P., & Herzog, M. H. (2008). Time-course and surround modulation of contrast masking in human vision. *Journal of Vision*, 8(3), 23–23.
- Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, 9(2), 5–5.
- Saarela, T. P., Westheimer, G., & Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *Journal of Vision*, 10(10), 17–17.
- Sayim, B., Greenwood, J. A., & Cavanagh, P. (2014). Foveal target repetitions reduce crowding. *Journal of Vision*, 14(6), 4–4.
- Sayim, B., Manassi, M., & Herzog, M. (2014). How color, regularity, and good Gestalt determine backward masking. *Journal of Vision*, 14(7), 8–8.
- Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Figural grouping affects contextual modulation in low level vision. *Journal of Vision*, 8(6), 436–436.

- Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychological Science*, 21(5), 641–644.
- Sayim, B., Westheimer, G., & Herzog, M. H. (2011). Quantifying target conspicuity in contextual modulation by visual search. *Journal of Vision*, 11(1), 6–6. <https://doi.org/10.1167/11.1.6>
- Schwiedrzik, C. M., Melloni, L., & Schurger, A. (2018). Mooney face stimuli for visual perception research. *PLoS One*, 13(7), e0200106.
- Scolari, M., Kohnen, A., Barton, B., & Awh, E. (2007). Spatial attention, preview, and popout: Which factors influence critical spacing in crowded displays? *Journal of Vision*, 7(2), 7–7.
- Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology*, 75(2), 221–242.
- Siderov, J., Waugh, S. J., & Bedell, H. E. (2013). Foveal contour interaction for low contrast acuity targets. *Vision Research*, 77, 10–13.
- Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5), 13–13.
- Sun, H.-M., & Balas, B. (2015). Face features and face configurations both contribute to visual crowding. *Attention, Perception, & Psychophysics*, 77(2), 508–519. <https://doi.org/10.3758/s13414-014-0786-0>
- Tannazzo, T., Kurylo, D. D., & Bukhari, F. (2014). Perceptual grouping across eccentricity. *Vision Research*, 103, 101–108. <https://doi.org/10.1016/j.visres.2014.08.011>
- Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research*, 51(11), 1273–1278.
- Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, 32(7), 1349–1357.
- Van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLoS Comput Biol*, 6(1), e1000646.
- Vickery, T. J., Shim, W. M., Chakravarthi, R., Jiang, Y. V., & Luedeman, R. (2009). Supercrowding: Weakly masking a target expands the range of crowding. *Journal of Vision*, 9(2), 12–12.
- Wallis, T., Funke, C., Ecker, A., Gatys, L., Wichmann, F., & Bethge, M. (2017). Towards matching peripheral appearance for arbitrary natural images using deep features. *Journal of Vision*, 17(10), 786–786.
- Waugh, S. J., & Formankiewicz, M. A. (2020). Grouping Effects on Foveal Spatial Interactions in Children. *Investigative Ophthalmology & Visual Science*, 61(5), 23–23. <https://doi.org/10.1167/iov.61.5.23>
- Westheimer, G., & Hauske, G. (1975). Temporal and spatial interference with vernier acuity. *Vision Research*, 15(10), 1137–1141.
- Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, 15(4), 160–168.
- Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *Josa a*, 14(9), 2057–2068.
- Wolford, G., & Chambers, L. (1983). Lateral masking as a function of spacing. *Perception & Psychophysics*, 33(2), 129–138. <https://doi.org/10.3758/BF03202830>
- Yeshurun, Y., & Rashal, E. (2010). Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision*, 10(10), 16–16.
- Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81(1), 141.
- Zhang, X., Huang, J., Yigit-Elliott, S., & Rosenholtz, R. (2015). Cube search, revisited. *Journal of Vision*, 15(3), 9–9.

Chapter 3: Crowding reveals fundamental differences in local vs. global processing in humans and machines

Doerig, A.[†], Bornet, A.[†], Choung, O. H., Herzog, M. H.

Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[†]These authors contributed equally to this work.

Abstract

Feedforward Convolutional Neural Networks (ffCNNs) have become state-of-the-art models both in computer vision and neuroscience. However, human-like performance of ffCNNs does not necessarily imply human-like computations. Previous studies have suggested that current ffCNNs do not make use of global shape information. However, it is currently unclear whether this reflects fundamental differences between ffCNN and human processing or is merely an artefact of how ffCNNs are trained. Here, we use visual crowding as a well-controlled, specific probe to test global shape computations. Our results provide evidence that ffCNNs cannot produce human-like global shape computations for principled architectural reasons. We lay out approaches that may address shortcomings of ffCNNs to provide better models of the human visual system.

Introduction

Vision is a complex process that remained beyond the reach of computer systems for decades. Only recently, deep feedforward Convolutional Neural Networks (ffCNNs) have shown tremendous success in an impressive number of computer vision tasks, ranging from object recognition (1) and segmentation (2), to image synthesis (3,4) and scene understanding (5). ffCNNs and the human visual system share several similarities. For example, after training on complex visual datasets such as ImageNet (6), ffCNN neural activities show high correlations with human and non-human primate neural activities (7–10) and the receptive fields of neurons in the earlier layers of these ffCNNs are qualitatively similar to those in the retina and early visual cortex (11,12). Because of these similarities, ffCNNs trained on complex visual tasks were proposed as models of the human visual system (7–9,13,14). However, human-like performance of ffCNNs does not necessarily imply human-like computations. Importantly, several studies have shown that ffCNNs usually rely on local features while humans strongly rely on global shape information (15–18).

There are two main options to explain why ffCNNs do not process global shape like humans. First, this difference may come from *training*. ffCNNs are typically trained on ImageNet. It is interesting and surprising that local features seem to be the easiest way for these networks to classify natural images. However, a different training set in which local features are not predictive of the classes may require networks to rely on global shape computations. To address this possibility, Geirhos et al. (19) created a new dataset in which textural information was of no avail for object recognition. They used a textural algorithm (20) to randomly swap textures in ImageNet. For example, the texture of a cat image was replaced by elephant-skin texture. This training dataset biased an ffCNN (ResNet50; 21) towards shape-level features, because textural information was no longer useful for classifying this dataset. They validated the network's shape-bias by showing increased robustness to local noise and textural changes.

Alternatively, ffCNNs may be incapable of matching human global computations for principled *architectural* reasons. Even though Geirhos et al.'s network was able to ignore local features, it may not use global computations in the same way as humans. One difficulty in addressing this question is that there is no consensus about how to experimentally diagnose *how* deep networks compute global information.

To specifically investigate local vs. global processing in humans and machines, we use visual crowding as an experimental probe. Crowding is the technical term for the everyday observation that objects are harder to perceive in clutter. Neighbouring visual elements are perceived as jumbled or indistinct, and are hard to recognize (Fig 1; 22–24). This phenomenon is strongest in the periphery, but also occurs in the fovea (25,26). This phenomenon is ubiquitous in natural vision since elements rarely appear in isolation (Fig 1a). Crowding can also be studied with high precision in psychophysical experiments. For example, when a vernier target (i.e., two vertical bars with a horizontal offset) is presented alone, the direction of the horizontal offset is easy to report. This task becomes harder in the presence of a surrounding square flanker (Fig 1b, column 1). Interestingly, the *global* configuration of flankers across the entire visual field determines crowding. For example, adding flankers as far away as 8.5 degrees from the 200 arcsec target can *improve* performance depending on the global configuration (*uncrowding*; Fig 1b; 27,28). This strong dependency of performance on global configurations provides a qualitative signature which can easily be tested in models. Importantly, (un)crowding occurs across multiple paradigms (26,29,30) and is not restricted to vision (31,32). Hence, (un)crowding is not an idiosyncratic effect related to a specific paradigm. It rather reflects a general strategy used by the brain. This kind of general strategy for vision is precisely what we expect models to explain.

Crowding effects have been shown in ffCNNs (17,33,34), and may occur by pooling the target and nearby flankers along the processing hierarchy. We hypothesize that this mechanism may not produce uncrowding because simple pooling can only deteriorate target-relevant information when flankers are added (Fig 1c). However, intuitions are not to be trusted in complex systems with millions of parameters. Furthermore, new global processing strategies may emerge in shape-biased networks such as Geirhos et al.'s. Hence, it is currently unclear whether ffCNNs can carry out human-like global computations that lead to (un)crowding.

Here, we thoroughly investigated (un)crowding in AlexNet (1), an ffCNN that was used as a model of the human visual system (7,12), ResNet50 (21), a more sophisticated ffCNN, and the shape-biased network by Geirhos et al. (19). We provide experimental evidence suggesting that it is the *architecture* of ffCNNs that prevents them from performing human-like global computations, and not the training procedure.

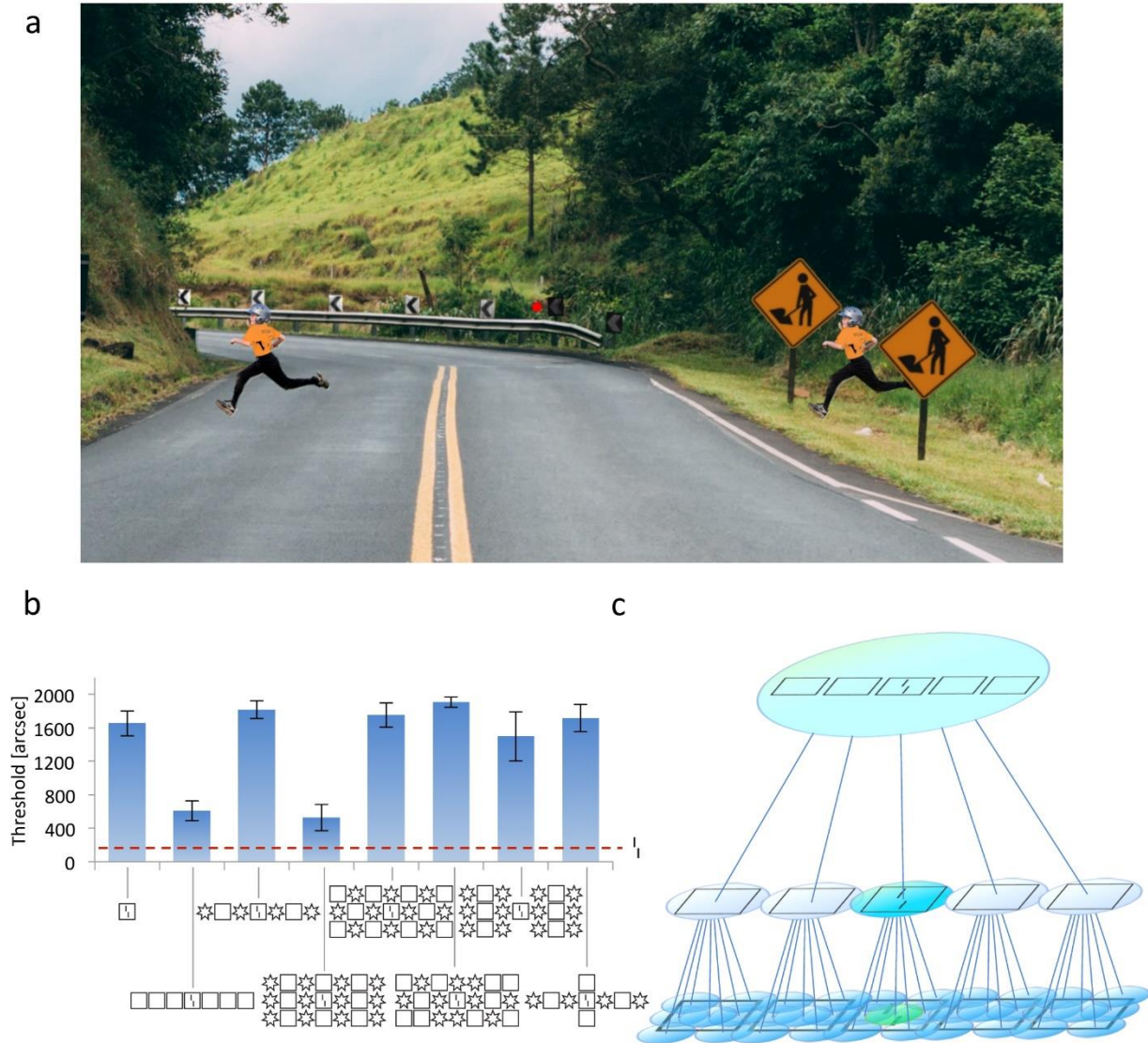


Fig 1. a. In crowding, perception of a target deteriorates in the presence of nearby visual elements. Crowding is ubiquitous in everyday vision, since elements rarely appear in isolation. When fixating on the central red dot, it is more difficult to spot the kid on the right than on the left, because of the nearby signposts. **b. (Un)crowding.** Visual elements can be rescued from crowding depending on the global configuration of flankers (*uncrowding*). In this experiment, observers reported the horizontal offset direction of two vertical bars (i.e., a *vernier*) presented at 9° of eccentricity. The vernier was presented either alone (red dashed line) or surrounded by a flanker configuration (x-axis). The y-axis shows the offset for which observers correctly report the vernier offset direction in 75% of the trials (threshold; performance is good when the threshold is low). When the vernier is presented alone, the task is easy (red dashed line). Adding a flanking square (column 1) makes the task much harder, a classic crowding effect. When more squares are added, performance recovers almost to the unflanked level (second column, *uncrowding*). Uncrowding strongly depends on the configuration (columns 2 to 8). For example, column 4 shows a configuration of flankers with a strong uncrowding effect. In comparison, column 5 has the same flankers but in a different configuration producing strong crowding. **c. Crowding in ffcNNs.** In the feedforward framework of vision, embodied by ffcNNs, crowding occurs by pooling of visual features across a hierarchy of local feature

detectors. In this example, a stimulus with five squares and a vernier target is presented. Each circle represents a neuron and shows the elements in its receptive field. In early layers, receptive fields are small and the vernier is in the receptive field of a single neuron (green). Neighboring neurons respond to parts of the squares (blue). At this level, the vernier is well represented. In the next layer, however, information about the vernier is pooled with information of the surrounding flanker. Vernier-related information is “corrupted” by the flankers, making the offset direction harder to decode (crowding; blue-green). In subsequent layers, even more target-unrelated information is pooled. For this reason, we hypothesize that adding more flankers may always lead to more crowding in ffCNNs. Modified from Doerig, Bornet et al. (17) with permission.

Methods

The code is available online at <https://github.com/adriendoerig/Doerig-Bornet-Choung-Herzog-2019>. The supplementary information for this Chapter is provided at this address.

Experiment 1a

We presented different (un)crowding stimuli to AlexNet (trained on ImageNet prior to our experiment) and assessed how information about the target vernier is preserved along the network hierarchy. We used decoders to detect vernier offset direction based on the activity in each layer (Fig 2). Each layer had its own decoder, consisting of batch normalization (35), followed by a hidden layer of 512 units, followed by an ELU non-linearity (36), finally projecting to a softmax layer composed of 2 nodes coding for left and right offsets. The weights of AlexNet were frozen during this process, only the decoder weights were trained. The decoders were trained using Adam optimizers (37) to minimize the cross-entropy between the predicted and the presented vernier offsets. Each image in the training set consisted of a vernier plus a non-overlapping random configuration of flankers (composed of 18x18 pixels squares, circles, hexagons, octagons, stars or diamonds). These configurations had between 1 and 7 columns and between 1 and 3 rows of flankers of the same shape. We added Gaussian noise to each image. Training was successful, i.e., the network was well able to detect the vernier offset direction in the training images.

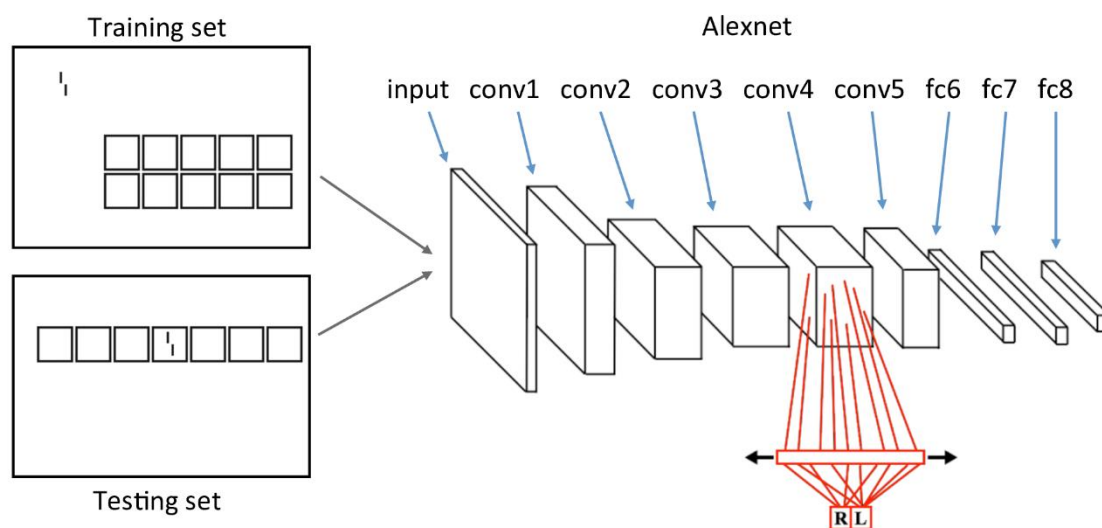


Fig 2. Different stimuli were fed to AlexNet. AlexNet's weights were trained on ImageNet prior to the experiment and were frozen during the experiment. To investigate how well information about the vernier offset is preserved

throughout the network hierarchy, we trained one decoder (in red) at each layer to discriminate the vernier offset direction based on the activity elicited by the stimulus in this layer. For example, the stimulus at the top left of this figure is presented. This elicits activities in each layer of AlexNet and the decoders are trained to retrieve the offset direction based on this activity. Only the decoders are trained (red). In the training set, the vernier and a flanker configuration were simultaneously shown, but never overlapped (top). In the testing set, we presented 72 different (un)crowding configurations and measured performance for each configuration and each layer. In these testing images, the vernier was always surrounded by the flanker configuration (bottom). In this example, configurations of squares are shown, but we also used different shapes (see main text).

Our main question was how the network generalizes to the (un)crowding stimuli. Importantly, during training, the vernier target and the flanking configurations were presented simultaneously but never overlapped (Fig 2). During testing the vernier was surrounded by different flanker configurations, as in the psychophysical (un)crowding stimuli (Fig 2). The testing set consisted of 72 different configurations of flankers with Gaussian noise. There were 6400 trials per configuration with the configuration presented at different locations. For each layer of AlexNet, performance was measured as the proportions of correct vernier offset discrimination made by the decoder. We repeated this entire procedure 5 times, including training and testing, and report averaged performances.

Experiment 1b

We tested an ffcNN with a more sophisticated architecture (ResNet50) trained on ImageNet, and the same ffcNN architecture trained on a dataset tailored to bias the network towards global shape computations (i.e., Geirhos et al.'s shape-biased version of ResNet50). To this end, we applied exactly the same procedure as in experiment 1a to both the original version of ResNet50 and Geirhos et al.'s shape-biased version. The only difference was that we used 64 hidden units instead of 512, because this achieved better performance (i.e., better classification performance on crowded conditions).

Experiment 2

In experiment 2, we investigated which parts of the stimulus configurations the network mainly relies on by using an occlusion sensitivity measure (similarly to 12). We used the networks with decoders trained in experiment 1. For a given configuration, we collected the vernier offset decoder's output at each layer. Then we slid a 6x6 pixels Gaussian noise patch over the entire configuration and measured for each patch position P and network layer L how much the noise

patch affected the vernier offset discrimination. The noise patch had the same statistics as the background noise, effectively removing parts of the stimulus. The rationale is that when the patch occludes parts of the stimulus, which are important for classification, decoder predictions should be strongly affected. On the other hand, if the patch occludes an unimportant part of the stimulus, decoder predictions should not be affected. Since the global stimulus configuration matters for uncrowding, we were interested to see if the network relies on the global configuration or if it simply focused on the region close to the vernier.

For each patch location P and layer L , we quantified how much the noise patch biased vernier offset classification towards or away from the correct response:

$$score_{P,L} = \frac{\{\vec{T} \cdot (\vec{y}_{P,L} - \vec{x}_L)\}_{left_vernier}}{2} + \frac{\{\vec{T} \cdot (\vec{y}_{P,L} - \vec{x}_L)\}_{right_vernier}}{2}$$

Where $\vec{x}_L = (x_1, x_2)_L$ is the output of the decoder for layer L on the original stimulus *without* a noise patch (x_1 and x_2 respectively correspond to the network's prediction for a left- or right-offset vernier), $\vec{y}_{P,L} = (y_1, y_2)_{P,L}$ is the output of the decoder for layer L *with* the noise patch at position P and \vec{T} is a vector equal to $(+1, -1)$ if the correct vernier offset is left and $(-1, +1)$ otherwise. To avoid biases related to offset direction, we computed the mean score of the left- and right-offset versions of each stimulus.

Using this procedure, we obtained maps indicating which regions of a stimulus are most important for vernier offset discrimination. We used four different stimuli from Manassi et al. (27): a vernier alone, a vernier flanked by one square (leading to crowding in humans), a vernier flanked by a row of seven squares (leading to uncrowding in humans), and a vernier flanked by a row of seven alternating squares and stars (no uncrowding in humans). Additional stimuli are shown in the supplementary material.

Results

Experiment 1a

Unlike humans, AlexNet shows crowding but *not uncrowding*. The vernier offset is easily decoded from each layer when the vernier is presented alone, and performance drops when a single flanker is added. Crucially, performance deteriorates further when more flankers are added, regardless of the shape type (Fig 3a). Squares produced more crowding than circles, hexagons, octagons or diamonds, presumably because the vertical bars of the squares interfered with the vernier more strongly. These results hold for all layers of AlexNet (supplementary material).

Fig 3b shows that, unlike humans who show strong uncrowding depending on the configuration, only the number of shapes seems to affect crowding in AlexNet – and not the configuration. Although certain configurations with three flankers have a higher percentage of correct response than certain configurations with a single flanker, this effect is driven by the shape type and not by the configuration of shapes. For example, the networks are better at dealing with diamonds than squares (Fig 3a; probably squares interfere more with verniers due to their vertical edges). Still, adding extra shapes always deteriorates performance compared to a single shape, regardless of the configuration. This pattern of results is similar in all layers of AlexNet (supplementary material).

Experiment 1b

We applied the same analysis to the original ResNet50 and Geirhos et al.'s shape-biased version of ResNet50. The results for both networks are qualitatively similar to the results for AlexNet in experiment 1a (Fig 3c&d). One difference is that the performance of the decoder is always below chance level with diamonds. This indicates that information about the vernier offset survives, even though the diamond flanker reverses the prediction. Adding additional diamond flankers brings performance closer to chance level, indicating that less information about the vernier offset survives, i.e., crowding increases when adding flankers. Another difference is that the squares lead to the least amount of crowding, contrary to AlexNet.

First, these results show that using a more sophisticated ffcnn (i.e., ResNet50) does not allow ffcnn to explain global uncrowding effects. Second, crucially, Geirhos et al.'s training method

to bias ffcNNs towards shape does not lead to uncrowding either. This suggests that ffcNNs do not carry out human-like shape level computations for *architectural* reasons, and not because of the way they are *trained*.

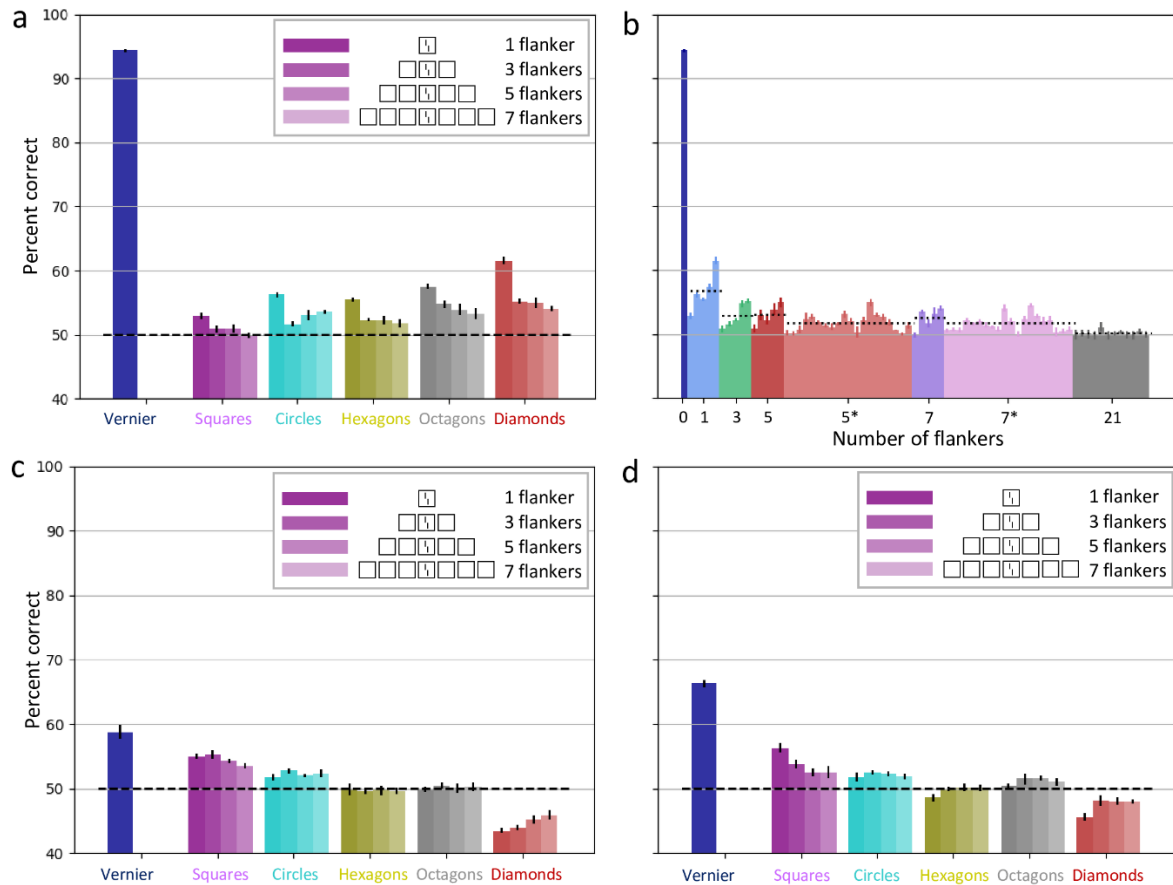


Fig 3. a. Vernier offset discrimination performance for AlexNet with an increasing number of identical flankers.

The x-axis shows different flanker configurations. Each color corresponds to one flanker shape, and brighter colors indicate more flankers (from darkest to lightest: 1, 3, 5 & 7 identical flankers). The single dark blue bar on the left corresponds to the vernier alone condition. The y-axis indicates the percentage of correct vernier offset responses. Unlike humans, for whom performance improves when more identical flankers are added (Fig 1b, columns 1&2; 27), performance deteriorates or stagnates for AlexNet with all flanker shapes. The results of this figure are decoded from layer 5 of AlexNet. Decoding vernier offsets from the other layers in AlexNet led to similar results (see supplementary material). **b. Vernier offset discrimination performance for AlexNet with 72 configurations.** The x-axis shows different flanker configurations sorted by number of flankers. Different colors correspond to different kinds of flanker configurations. The labels correspond to the number of flankers in the configuration, and an asterisk indicates alternating shapes (e.g. square-circle-square-circle-square). From left to right: vernier alone, single flanker, 3 identical flankers, 5 identical flankers, 5 flankers alternating between two shapes, 7 identical flankers, 7 flankers alternating between two shapes and configurations of 3x7 flankers. The y-axis indicates percent correct of vernier offset discrimination for each flanker configuration (the dashed lines shows the mean percent correct for each kind of flanker configuration). The results of this figure are decoded

from layer 5 of AlexNet. Decoding vernier offsets from the other layers in AlexNet led to similar results (see supplementary material). **c&d. Vernier offset discrimination performance with an increasing number of identical flankers for ResNet50 (original version in c, Geirhos et al.'s shape-biased version in c).** The results for both networks are qualitatively similar to the results for AlexNet in panel a. The results of this figure are decoded from the output of the third bottleneck unit (see our shared code and He et al. 21). Decoding vernier offsets from the other layers led to similar results (see supplementary material).

Experiment 2

Uncrowding requires global computations across large regions of the visual space. The configuration in its entirety determines performance and not only the elements in the neighborhood of the target (17,27,28). As mentioned, it has been proposed that ffCNNs focus largely on local features. This is indeed what we observed in experiment 2 in AlexNet (Fig 4), ResNet50 (supplementary material), and Geirhos et al.'s shape-biased version of ResNet50 (Fig 4): only elements in a local region around the target matter for classification. The same results also hold for the eight other stimulus types we tested (supplementary material). In general, as expected, occluding the vernier target deteriorates performance and occluding parts of the flanker surrounding the vernier improves performance. Occluding other parts of the stimulus, however, does not generally affect performance. Certain cases are harder to explain, such as the 1square condition shown in the top right panel of Fig 4, in which occluding parts of the vernier improved classification. Although we cannot provide a definitive explanation, we suggest that this may be due to the classifier confusing a vertical bar of the square with a vertical vernier bar. Alternatively, this may be due to the background noise present in each stimulus. In rare cases, the occluder has an effect even when it does not cover the stimulus (e.g. in the bottom right panel of Fig 4). These cases are also probably due to background noise. Aside from these small peculiarities, the finding that only elements in the neighborhood of the vernier affect classification is very stable over all stimuli and network layers (see images and animations in the supplementary material).

These results suggest that the inability of ffCNNs to explain uncrowding stems from their focus only on local features close to the vernier. Importantly, although Geirhos et al.'s shape-biased network is biased towards global features, still, performance seems determined only by elements close to the vernier.

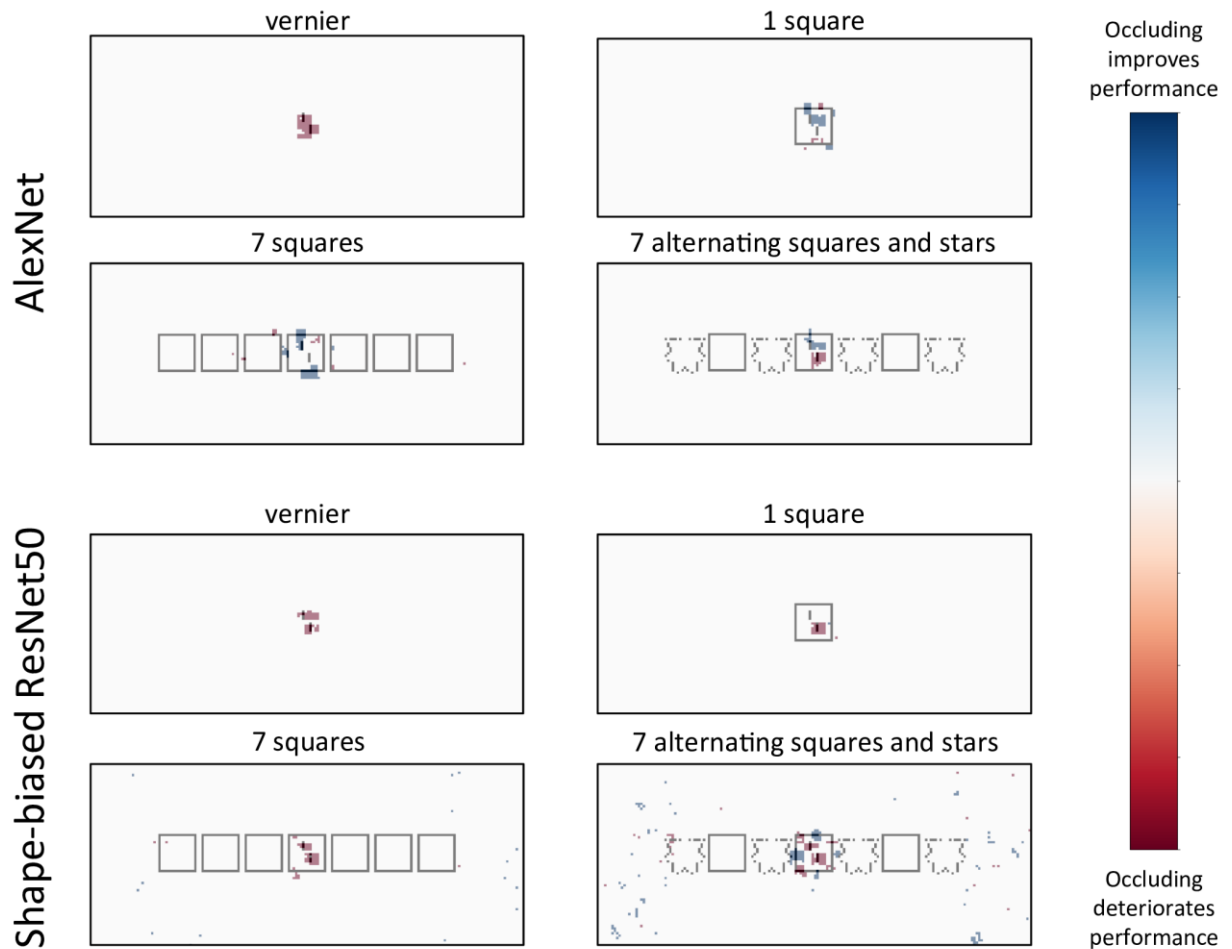


Fig 4: Occlusion analysis. Results of the occlusion analysis for AlexNet (*top*) and the shape biased ResNet50 (*bottom*). Stimuli on the left lead to good performance in humans, while stimuli on the right lead to strong crowding in humans (27). For both AlexNet and the shape biased ResNet50, the network’s decisions rely only on local elements in the target neighborhood regardless of the global stimulus configurations. To create these maps, we summed the maps for each layer of Alexnet to show which stimulus regions are most relevant across the network. For the shape-biased ResNet50, we used the third convolutional layer in the first bottleneck, and the output of the first 9 bottleneck units (see our shared code and He et al., 2016). We then applied a threshold to each map at 0.4 times the maximal value in the map, for visibility. Per-layer results without thresholding can be found in the supplementary material, as well as animations showing what happens as the threshold value is changed. Results for the original ResNet50 and other layers of the shape-biased network are also shown in the supplementary material.

Discussion

(Un)crowding is ubiquitous. It occurs in vision, audition and haptics (24,27,31,32). This pervasiveness is not surprising because elements rarely appear in isolation. Any perceptual system needs to cope with crowding to process information in cluttered environments. (Un)crowding is a probe into how the visual system computes global information.

In this contribution, we asked whether large ffCNNs trained on complex visual tasks can explain (un)crowding. We chose this approach because these ffCNNs are often used as brain models. The idea is that the weights learned by these ffCNNs to solve complex visual tasks may lead to human-like visual processing. For this reason, we did not change the ffCNN weights for quantifying (un)crowding, i.e., we only trained the additional decoders. We found that these ffCNNs do not seem to carry out human-like global computations.

Experiment 1 shows that current ffCNNs do not explain (un)crowding. In other words, training an ffCNN on a complex natural image recognition task does not automatically yield a network performing similarly to the human visual system. Experiment 2 suggests that this is due to the inability of ffCNNs to take the entire stimulus configuration into account. In ffCNNs, only elements in the target's neighborhood affect performance. Global features do not affect how local parts are processed. In humans, on the other hand, the global configuration strongly affects processing of local parts. For example, vernier offset information can be "rescued" by certain global configurations.

This difference could not be remedied by a different *training* protocol. Indeed, all our results also hold for Geirhos et al.'s shape-biased ffCNN. We suggest that, although Geirhos et al.'s training procedure successfully biased the networks towards global features, it does not show human-like global shape computations. Indeed, the network still seems limited to combining features by pooling along the feedforward cascade. Hence, unlike in humans, global configuration cannot affect processing of local parts. For these reasons, our results suggest that the inability of ffCNNs to perform human-like object shape processing is rooted in their feedforward pooling *architecture*. Because of this pooling, performance deteriorates when flankers are added. For this principled reason, we propose that ffCNNs cannot produce uncrowding in general, independently of the specific ffCNN, training procedure and loss function. In support of this proposal, we showed in a separate contribution that ffCNNs

specifically trained on classifying verniers and flanking shapes, as well as counting the number of flankers, do not produce global (un)crowding either (38).

Global processing is not only an issue for ffCNNs but for other models too. We showed that no existing model of crowding based on local and feedforward computations can explain uncrowding (17,27,30,39). There seems to be a principled difference in computational strategies, based on architecture, between humans and feedforward pooling systems.

Hence, despite their well-known power, further aspects need to be incorporated into ffCNNs. We propose that recurrent, global grouping and segmentation is crucial to explain how the brain deals with global configurations (17,38). Specifically, we propose that a flexible recurrent grouping process determines which elements are grouped into an object. In the case of (un)crowding, elements are first grouped together and then only elements within a group interfere with each other. If the configuration of flankers ungroups from the target, the target is released from crowding. Francis, Manassi, and Herzog (40) proposed a spiking neural network with a dedicated recurrent grouping process, which is able to explain why (un)crowding occurs (see also Bornet et al.; 41). However, this model is tailored to group oriented edges and cannot generalize to grouping of more complex features. Deep learning models are promising because they are more flexible and can be trained to deal with any kind of stimulus.

Doerig et al. (38) showed that capsules networks (42), combining CNNs with a recurrent grouping and segmentation process, can explain (un)crowding, including temporal characteristics of uncrowding. Linsley et al. (43) proposed recurrent grouping and segmentation modules to improve CNNs, and there are several other approaches to experiment with grouping and segmentation in recurrent network architectures (8,44–46). More work is needed to compare and characterize computations in different recurrent architectures.

Our results contribute to the expanding literature showing that there is much more to vision than combining local feature detectors in a feedforward hierarchical manner (15–17,38,42,43,45–53). In line with the present findings, many studies have highlighted other fundamental differences between ffCNNs and humans in local vs. global processing. For example, Baker et al. (15) showed that ffCNNs but not humans are affected by local changes

to edges and textures of objects. Brendel and Bethge (16) showed that ffCNNs classify ImageNet images almost as well when using small local image patches than when using the entire images. These results clearly show that image classification is underconstrained as a testbed. For this reason, well-controlled psychophysical stimuli, which allow detailed analysis, should be used in addition to image classification (54). Simply testing whether deep learning systems reproduce idiosyncratic illusions, without linking them to computational mechanisms, does not provide principled insights. Hence, an important question will be what are the crucial benchmarks targeting principled computational processes. Here, using crowding, we showed a fundamental difference in local vs. global processing between humans and ffCNNs, and suggest that grouping and segmentation are promising additions to make deep neural networks better models of vision.

Historically, psychophysical results were seen as stepping stones towards object recognition models. Today, the picture has been reversed: we have powerful artificial vision models, but they do not reproduce even simple psychophysical results. The fact that ffCNNs can solve complex visual tasks in a different way than humans reveals that there are many ways of doing so. There are many roads to Rome. Despite the diversity of possible strategies to solve complex vision tasks, deep insights can be derived by comparing the crucial underlying computations adopted by different systems.

References

1. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
2. Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K. Detectron. 2018.
3. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: *Advances in neural information processing systems*. 2014. p. 2672–80.
4. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. *ArXiv Prepr ArXiv181204948*. 2018;
5. 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.
6. 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.
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. Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, et al. Task-Driven Convolutional Recurrent Models of the Visual System. *ArXiv Prepr ArXiv180700053*. 2018;
9. 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.
10. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 586–95.
11. 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;
12. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Springer; 2014. p. 818–33.
13. Kietzmann TC, McClure P, Kriegeskorte N. Deep neural networks in computational neuroscience. *bioRxiv*. 2018;133504.
14. VanRullen R. Perception science in the age of deep neural networks. *Front Psychol*. 2017;8:142.
15. Baker N, Lu H, Erlikhman G, Kellman PJ. Deep convolutional networks do not classify based on global object shape. *PLoS Comput Biol*. 2018;14(12):e1006613.
16. Brendel W, Bethge M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *ArXiv Prepr ArXiv190400760*. 2019;
17. 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 May 10;15(5):e1006580.
18. Kim T, Bair W, Pasupathy A. Neural coding for shape and texture in macaque area V4. *J Neurosci*. 2019;39(24):4760–74.
19. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Prepr ArXiv181112231*. 2018;
20. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 2414–23.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770–8.
22. Herzog MH, Sayim B, Chicherov V, Manassi M. Crowding, grouping, and object recognition: A matter of appearance. *J Vis*. 2015;15(6):5–5.
23. Levi DM. Visual crowding. *Curr Biol*. 2011;21(18):R678–9.
24. Whitney D, Levi DM. Visual crowding: a fundamental limit on conscious perception and object recognition. *Trends Cogn Sci*. 2011 Apr 1;15(4):160–8.

25. Malania M, Herzog MH, Westheimer G. Grouping of contextual elements that affect vernier thresholds. *J Vis.* 2007;7(2):1–1.
26. Sayim B, Westheimer G, Herzog MH. Gestalt factors modulate basic spatial vision. *Psychol Sci.* 2010;21(5):641–4.
27. Manassi M, Lonchamp S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis.* 2016 Feb 1;16(3):35–35.
28. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis.* 2012 Sep 1;12(10):13–13.
29. Herzog MH, Fahle M. Effects of grouping in contextual modulation. *Nature.* 2002 Jan;415(6870):433.
30. Pachai MV, Doerig AC, Herzog MH. How best to unify crowding? *Curr Biol.* 2016 May 9;26(9):R352–3.
31. Oberfeld D, Stahn P. Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PloS One.* 2012;7(10):e48054.
32. Overvliet KE, Sayim B. Perceptual grouping determines haptic contextual modulation. *Vision Res.* 2016 Sep 1;126(Supplement C):52–8.
33. Lonnqvist B, Clarke AD, Chakravarthi R. Object Recognition in Deep Convolutional Neural Networks is Fundamentally Different to That in Humans. *ArXiv Prepr ArXiv190300258.* 2019;
34. Volokitin A, Roig G, Poggio TA. Do deep neural networks suffer from crowding? In: *Advances in Neural Information Processing Systems.* 2017. p. 5628–38.
35. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv Prepr ArXiv150203167.* 2015;
36. Clevert D-A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). *ArXiv Prepr ArXiv151107289.* 2015;
37. Kingma DP, Ba J. Adam: A method for stochastic optimization. *ArXiv Prepr ArXiv14126980.* 2014;
38. Doerig A, Schmittwilken L, Sayim B, Manassi M, Herzog MH. Capsule Networks as Recurrent Models of Grouping and Segmentation. *bioRxiv.* 2019 Jan 1;747394.
39. Herzog MH, Manassi M. Uncorking the bottleneck of crowding: A fresh look at object recognition. *Curr Opin Behav Sci.* 2015;1:86–93.
40. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychol Rev.* 2017;124(4):483.
41. Bornet A, Kaiser J, Kroner A, Falotico E, Ambrosano A, Cantero K, et al. Running large-scale simulations on the Neurorobotics Platform to understand vision-the case of visual crowding. *Front Neurobotics.* 2019;13:33.
42. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: *Advances in neural information processing systems.* 2017. p. 3856–66.
43. Linsley D, Kim J, Serre T. Sample-efficient image segmentation through recurrence. *ArXiv181111356 Cs [Internet].* 2018 Nov 27 [cited 2019 Jun 27]; Available from: <http://arxiv.org/abs/1811.11356>
44. Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning. *ArXiv Prepr ArXiv160508104.* 2016;
45. Spoerer CJ, Kietzmann TC, Kriegeskorte N. Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. *bioRxiv.* 2019;677237.
46. Spoerer CJ, McClure P, Kriegeskorte N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front Psychol.* 2017;8:1551.
47. Funke CM, Borowski J, Wallis TSA, Brendel W, Ecker AS, Bethge M. Comparing the ability of humans and DNNs to recognise closed contours in cluttered images. In: *18th Annual Meeting of the Vision Sciences Society (VSS 2018).* 2018. p. 213.
48. Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. *Nat Neurosci.* 2019;22(6):974.
49. Kietzmann TC, Spoerer CJ, Sörensen L, Cichy RM, Hauk O, Kriegeskorte N. Recurrence required to capture the dynamic computations of the human ventral visual stream. *ArXiv Prepr ArXiv190305946.* 2019;
50. Kim J, Linsley D, Thakkar K, Serre T. Disentangling neural mechanisms for perceptual grouping. *ArXiv Prepr ArXiv190601558.* 2019;

51. Lamme VA, Roelfsema PR. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* 2000;23(11):571–9.
52. Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Caro JO, et al. Recurrent computations for visual pattern completion. *Proc Natl Acad Sci.* 2018;115(35):8835–40.
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.
54. RichardWebster B, Anthony S, Scheirer W. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Trans Pattern Anal Mach Intell.* 2018;

Chapter 4: Shrinking Bouma's window - Models of crowding in dense displays

Bornet, A.^{1†}, Doerig, A.^{1,2†}, Herzog, M. H.¹, Francis, G.³, Van der Burg, E.^{4,5}

¹Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne, Switzerland

²Artificial Cognitive Systems Group, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

³Department of Psychological Sciences, Purdue University, USA

⁴TNO, Netherlands Organization for Applied Scientific Research, Soesterberg, The Netherlands

⁵Brain and Cognition, University of Amsterdam, The Netherlands

[†]These authors contributed equally to this work.

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 pooling 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 et al. (1) 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. From previous studies, we know that visual grouping is a promising addition to explain why elements beyond Bouma's window interfere with the target. Here, we tested whether this explanation also helps explain crowding in dense displays. We compared the performance of models that include a grouping stage to models that do not. 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 that traditional models can be improved by adding a grouping stage.

Author summary

To understand human vision, psychophysical research has focused on very simple paradigms. Based on this research, vision was described as a cascade of feed-forward computations in which local features detectors pool information along the processing hierarchy to form complex and abstract features. However, recent data that uses more complex paradigms has challenged this view. For example, Van der Burg et al. (1) studied visual crowding in dense displays and found that the range at which visual elements interact with each other (which was believed to be half the eccentricity) is shrunk to the nearest neighbour distance only. In our study, we aim at understanding this discrepancy. From previous studies, we know that visual grouping is a promising addition to current models of vision. We compared the performance of different models of vision to the human data of Van der Burg et al. (2019). 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 explain naturally and consistently 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 (2–5). These basic features are pooled along the visual hierarchy to form more complex feature detectors, until neurons respond to objects (6–10). A strength of modelling visual perception as 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 (11–14).

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) but adding a square around the vernier severely impairs performance (i.e., crowding, Fig 1b, first column).

In the classic framework, such impairments are explained by flankers and target features being pooled along the visual hierarchy (15–18). 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, affect each other (19–22) and b) adding more flankers within this window always leads to more crowding because more irrelevant information is pooled. Bouma's window is approximately equal to half the target eccentricity.

However, recent research has shown many effects that cannot be explained in this framework. For example, flankers far from the target can in fact strongly *improve* performance, depending on the global configuration of the stimulus (uncrowding; Fig 1b, second to last columns; 11,12,23–29). As another example, it has been shown that detailed information can survive crowding (30,31). Hence, a) interactions are *not* restricted to Bouma's window and b) adding flankers does *not* always deteriorate information.

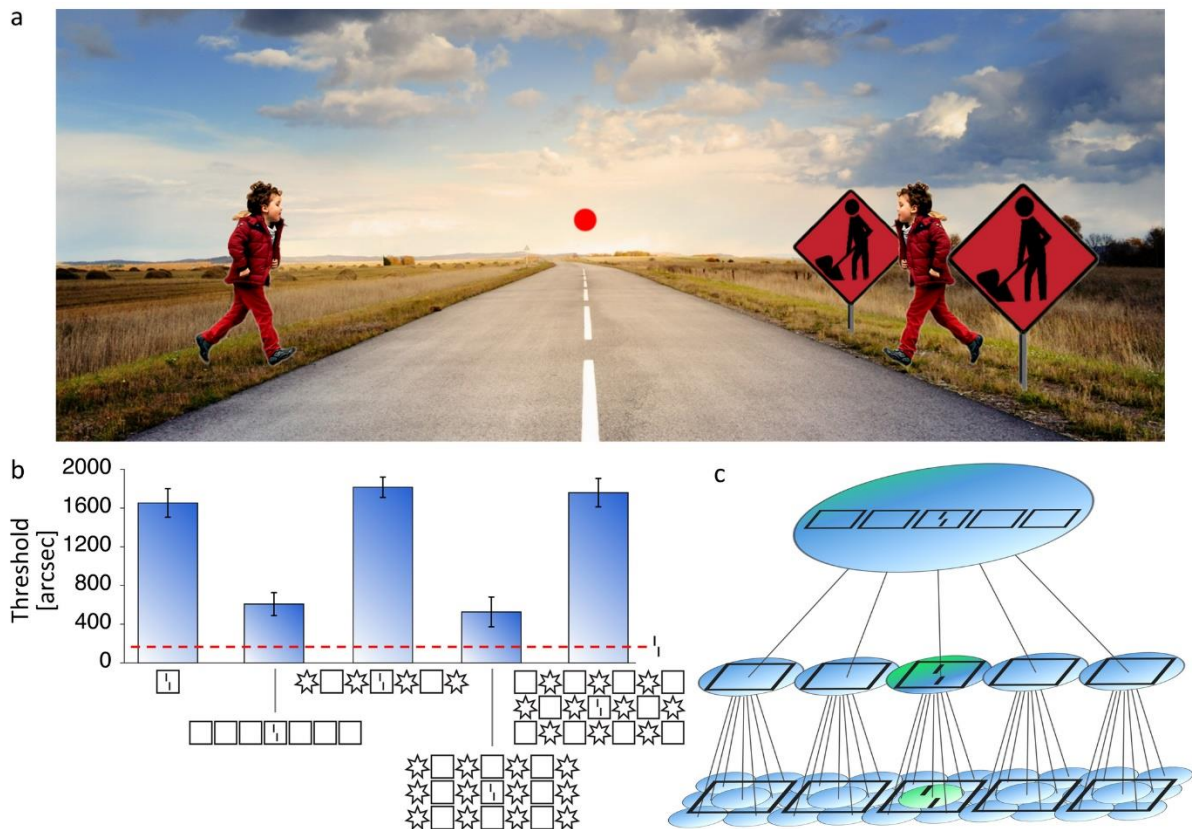


Fig 1. a. Visual crowding in everyday life. When looking at the red fixation dot, the child on the right is more difficult to identify than the same child on the left, because the nearby signposts lead to crowding (see also 32). **b.** Manassi et al. (33) presented a vernier in the periphery, surrounded by different flanker configurations. The y-axis shows the vernier offset threshold for 75% of correct responses (bigger numbers indicate a more difficult condition). 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 (34).

Obviously, studies with sparse displays cannot reveal these important effects. However, one of the main problems studying crowding in displays that contain a large number of flankers is that the configuration space increases exponentially with the number of flankers. For example, a relatively simple array of 8 by 8 either vertical or horizontal flanking bars has more possible configurations than there are seconds since the Big Bang. Among all these possible

configurations, it is unknown how many may show interesting effects that are not captured by the classic framework of vision. How can these configurations be discovered?

Recently, Van der Burg et al. (1) 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 showed 15-by-19 arrays (284 distractors and 1 target), containing either vertical or horizontal flankers at every position (*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 the problem of combinatorial explosion, Van der Burg et al. (1) used a genetic algorithm (GA; Fig 2b, bottom; 35). In this study, participants performed the orientation discrimination task. Subsequently, for each participant, the displays that led to the highest accuracy were selected and combined using a 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 36–38 for a similar methodology to study visual search in complex displays). Using this procedure, performance increased dramatically over generations (Fig 2c, bottom). Interestingly, this performance improvement was predominantly caused by changes to the target's nearest neighbours and, to a lesser extent, by other flankers within a radius of 1° (Fig 2c, top), which is in contradiction with Bouma's law. It seems as if Bouma's window has shrunk.

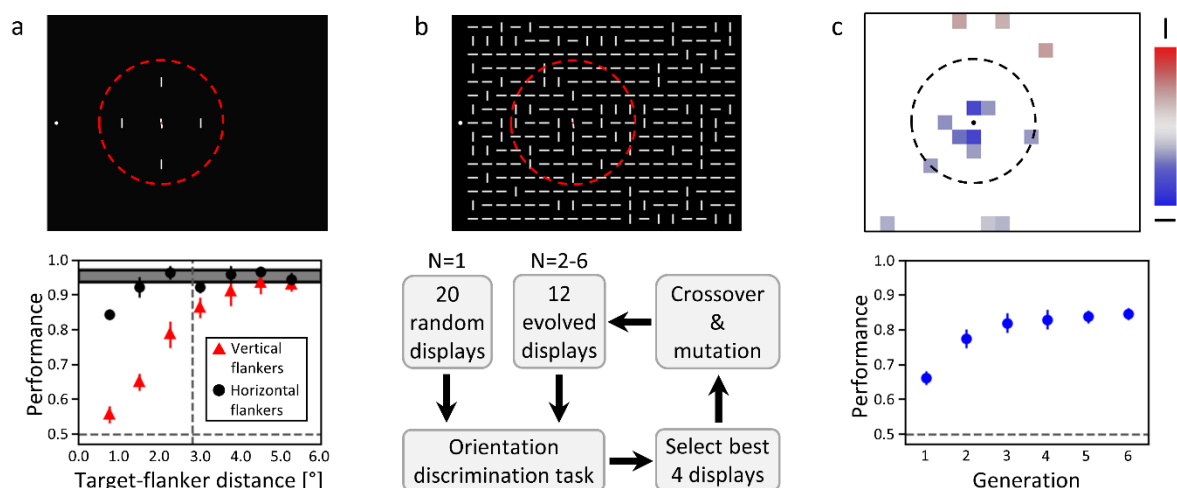


Fig 2. a. Top. Example display of the crowding experiment involving sparse displays in Van der Burg et al. (1). Observers reported whether the target was tilted to the left or right from vertical while fixating the white dot on the left. The target was surrounded by either four horizontal or vertical 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. (1). 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. (1). 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). More details are given in the Methods.

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. (1), 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, a population coding model (39), a model based on summary statistics (40) and a convolutional neural network classifier (34,41).

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 (34,42,43). For this reason, we also tested several models of crowding that include grouping and segmentation processes: a model of low-level segmentation (44), a convolutional neural network augmented with grouping processes (43,45) and a model that combined the population coding and the segmentation models. We compared the results obtained with both classes of models.

We show that only the models that contain a dedicated grouping mechanism explain the results of Van der Burg et al. (1). 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 (42), 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

The stimuli and the GA procedures were the same as in Van der Burg et al. (1). 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 2a, top, for an example display). 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 if it was tilted to the left or to the right from vertical. As in the human experiments of Van der Burg et al. (1), 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 generation. To reduce noise, the whole GA was run 4 times, like in Van der Burg et al. (1), where each participant performed 4 sessions.

For each model tested with the GA procedure, 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. (1). 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 proportion map where a red or a blue slot indicates that a vertical or a horizontal flanker was significantly preferred 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. (1), the statistical tests were not corrected for multiple comparisons to maximize the possibility of finding evidence for Bouma's law in the results. Colour intensity represents effect size. Black spaces indicate that neither vertical nor horizontal flankers were significantly preferred and therefore that a flanker at that location did not interfere with the target in dense displays. We call this the *preference 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 along the 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 preference measures as the *GA measures*.

In the GA procedure reported by Van der Burg et al. (1), the proportion of vertical flankers in the first generation of dense displays was set to lead to an initial performance of 67% for each individual 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. (1), i.e., 30% of vertical flankers in the first generation. If any model was far from this 67% requirement, we adapted the target orientation amplitude in dense displays, mentioning it in the description of the model.

Prior to the GA procedure, we fitted the tuneable parameters of each model using two control experiments. The goal was to find the best parameters for an optimal GA procedure and to have the fairest comparison between models. First, we measured model performance for the same sparse display experiment as in Van der Burg et al. (1). We call this the *sparse display measure* (see Fig 2a, bottom, for corresponding human results). Second, we measured model performance 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 no data was available for this variable (for more details, see [Suppl. Inf. C; SC8](#)).

The four measures (preference, performance, sparse display, proportion) are reported in the Results section by running each model 10 times, to simulate 10 different human subjects. The reported standard deviations are computed over these 10 runs. The code for the entire procedure is available at https://bitbucket.org/albornet/shrinking_boumas_window. Note that the main measure that was used to compare how well model behaviour reproduced human behaviour is the preference measure. All other measures were used as controls to ensure that the parameters of the model allowed convergence of the GA procedure. Also note that no quantitative assessment was performed to compare models in the preference measure, since the results were unequivocal. Either the model results fitted the human behaviour or did not.

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 of this Chapter ([Suppl. Inf. C](#)).

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. (1) 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 and the sparse display measures but failed to reproduce the human GA measures (Fig 3, 2nd 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, feedforward pooling framework of vision. First, we used a model based on the population coding idea (*Popcode model*; 39). 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*; 40), based on low-level summary statistics, which can be seen as high-dimensional pooling (17). Texture models may be particularly well suited for dense displays, because they encode complex natural information in a very efficient way. Third, we used a deep convolutional neural network (*CNN classifier*; 45). 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 in [Suppl. Inf. C; SC1-SC4](#).

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*; 45). 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 (i.e., it did not propose an actual interference mechanism), and the Capsule network was difficult to train properly, since only 1 network out of 10 trained networks could reach sufficient performance to be used in the GA procedure (for details, see [Suppl. Inf. C; SC5-SC6](#)). 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 an association would lead to a happy marriage between both families of crowding models (*Popart model*). Indeed, this combined model was able to reproduce human behaviour in the preference measure by proposing an actual interference mechanism, i.e., the one of the population coding model, and without requiring any training or pruning (Fig 3, last row; for more details, see [Suppl. Inf. C; SC7](#)).

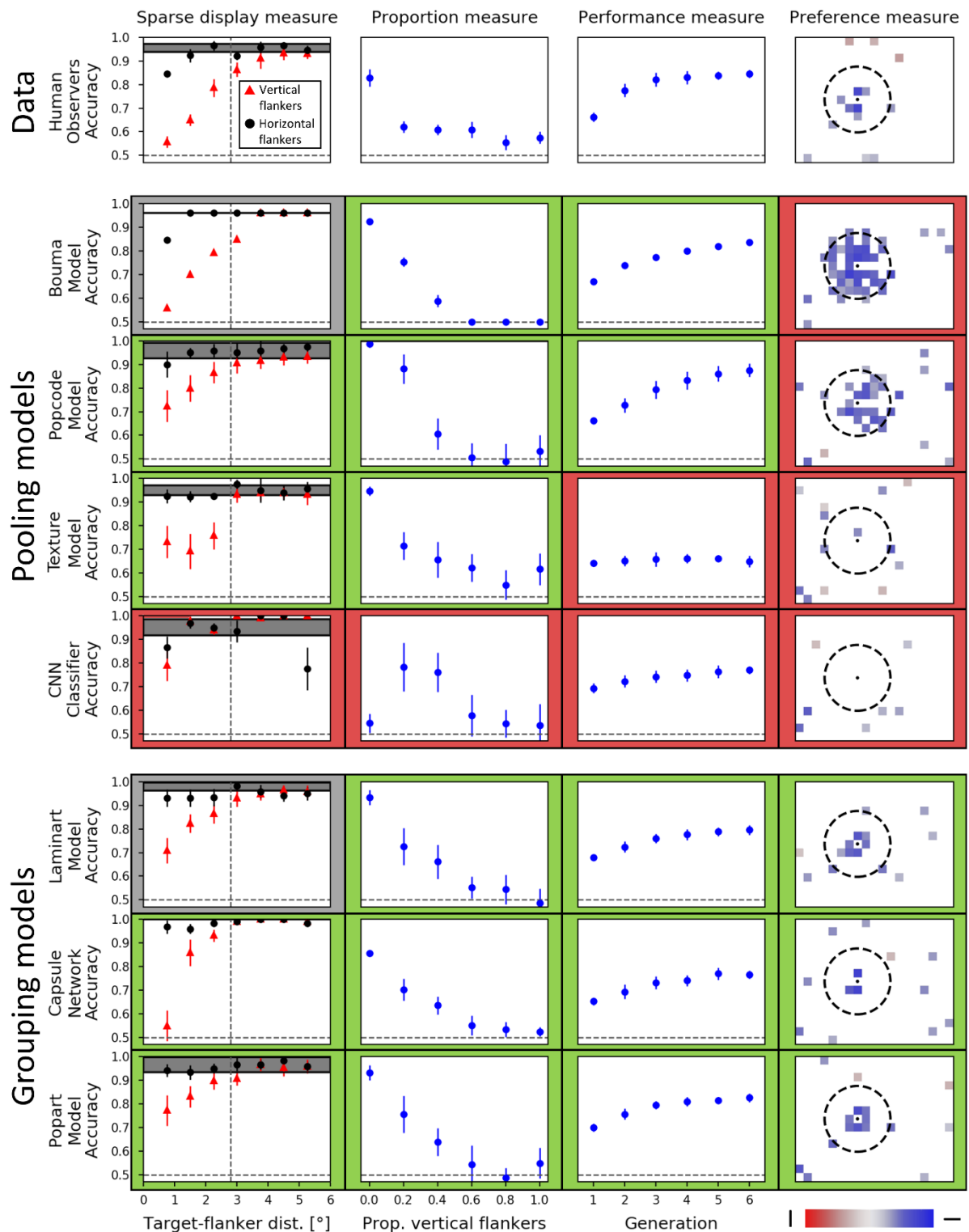


Fig 3. Results for all models, for the four measures described in the Methods section. The first row contains the human data. For every measure and every model, green/red frames indicate whether a model did or did not qualitatively reproduce the corresponding human data (for the performance measure, green corresponds to an improvement of at least 10 points of accuracy during the GA procedure), respectively. For the sparse display measure, a grey background indicates that the model results were fitted to the human data of Van der Burg et al. (1), which does not represent a true achievement of the model. The vertical dashed lines in the sparse display measure and the dashed circles in the preference 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 proportion 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.

Discussion

To understand crowding and vision in general, simple paradigms 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 simple paradigms and characterized it in detail as a local interference mechanism (15–18). However, 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 (1,33,46,47). Complex displays come with their own problems and questions, 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. (1) 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.

Here, we applied this 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 rule out or give support to general principles about human 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. (1). These models produced results in which either no element or too many elements within Bouma's window were highlighted by the GA procedure. In contrast, all the 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, combining a global grouping stage and a local interference stage led to the best results (Popart model).

Along the same line, Manassi et al. (33) 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 24,25). A similar extensive

comparison of models showed, once again, that only models that could reproduce these results contained a dedicated grouping stage (46; see also 34,43,49). 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. For all these reasons, it becomes clear that grouping, and not Bouma's window, determines which elements interfere with each other in human vision.

It is important to note that, contrary to our previous work (34,42), 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 which 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. Grouping and segmentation seem crucial to explain crowding in general (42,44,48,50). Moreover, it is known that texture models and other models based on pooling do not reproduce human grouping and segmentation (34,42,43,51,52). 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 grouping mechanisms. 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. (1), 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 for 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. (53) proposed another general recurrent grouping mechanism that is scalable to solve complex visual tasks at a state of the art level.

An important question for future research will be to 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 (51,52), naturalistic image segmentation (53) or spatiotemporal grouping and segmentation (54). 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.

References

1. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. *J Exp Psychol Hum Percept Perform.* 2017;43(4):690.
2. Gattass R, Gross CG, Sandell JH. Visual topography of V2 in the macaque. *J Comp Neurol.* 1981;201(4):519-39.
3. Gattass R, Sousa AP, Gross CG. Visuotopic organization and extent of V3 and V4 of the macaque. *J Neurosci.* 1988;8(6):1831-45.
4. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol.* 1962;160(1):106.
5. 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.
6. 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.
7. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci.* 1999;2(11):1019-25.
8. 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.
9. Ungerleider LG, Haxby JV. 'What' and 'where' in the human brain. *Curr Opin Neurobiol.* 1994;4(2):157-65.
10. Wallis G, Rolls ET. Invariant face and object recognition in the visual system. *Prog Neurobiol.* 1997;51(2):167-94.
11. Herzog MH, Sayim B, Chicherov V, Manassi M. Crowding, grouping, and object recognition: A matter of appearance. *J Vis.* 2015;15(6):5-5.
12. 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.
13. Herzog MH, Clarke AM. Why vision is not both hierarchical and feedforward. *Front Comput Neurosci.* 2014;8:135.
14. Saarela TP, Westheimer G, Herzog MH. The effect of spacing regularity on visual crowding. *J Vis.* 2010;10(10):17-17.
15. 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.
16. Pelli DG, Tillman KA. The uncrowded window of object recognition. *Nat Neurosci.* 2008;11(10):1129-35.
17. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. *J Vis.* 2019;19(7):15-15.
18. Wilson HR, Wilkinson F, Asaad W. Concentric orientation summation in human form vision. *Vision Res.* 1997;37(17):2325-30.
19. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Res.* 1973;13(4):767-82.
20. Levi DM. Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Res.* 2008;48(5):635-54.
21. Pelli DG, Palomares M, Majaj NJ. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *J Vis.* 2004;4(12):12-12.
22. 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.
23. Livne T, Sagi D. Configuration influence on crowding. *J Vis.* 2007;7(2):4-4.
24. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis.* 2012;12(10):13-13.
25. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. *J Vis.* 2013;13(13):10-10.

26. Manassi M, Lonchamp S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis.* 2016;16(3):35-35.
27. Pöder E. Crowding, feature integration, and two kinds of "attention". *J Vis.* 2006;6(2):7-7.
28. Saarela TP, Sayim B, Westheimer G, Herzog MH. Global stimulus configuration modulates crowding. *J Vis.* 2009;9(2):5-5.
29. Saarela TP, Herzog MH. Time-course and surround modulation of contrast masking in human vision. *J Vis.* 2008;8(3):23-23.
30. Manassi M, Whitney D. Multi-level crowding and the paradox of object recognition in clutter. *Curr Biol.* 2018;28(3):R127-33.
31. Whitney D, Haberman J, Sweeny TD. 49 From Textures to Crowds: Multiple Levels of Summary Statistical Perception. 2014;
32. Whitney D, Levi DM. Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends Cogn Sci.* 2011;15(4):160-8.
33. Manassi M, Lonchamp S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis.* 2016;16(3).
34. 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.
35. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press; 1992.
36. 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.
37. Kong G, Alais D, Van der Burg E. Competing distractors facilitate visual search in heterogeneous displays. *PLoS One.* 2016;11(8):e0160914.
38. Van de Weijert M, Van der Burg E, Donk M. Attentional guidance varies with display density. *Vision Res.* 2019;164:1-11.
39. 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.
40. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis.* 2009;9(12):13-13.
41. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems.* 2012. p. 1097-105.
42. 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.
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.
44. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychol Rev.* 2017;124(4):483.
45. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: *Advances in neural information processing systems.* 2017. p. 3856-66.
46. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. *J Vis.* 8 nov 2013;13(13):10.
47. 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.
48. Bornet A, Kaiser J, Kroner A, Falotico E, Ambrosano A, Cantero K, et al. Running large-scale 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. Herzog MH, Sayim B, Chicherov V, Manassi M. Crowding, grouping, and object recognition: A matter of appearance. *J Vis.* 2015;15(6):5-5.

51. Herrera-Esposito D, Coen-Cagli R, Gomez-Sena L. Flexible contextual modulation of naturalistic texture perception in peripheral vision. *bioRxiv*. 2020;
52. 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.
53. Linsley D, Kim J, Serre T. Sample-efficient image segmentation through recurrence. *ArXiv Prepr ArXiv181111356*. 2018;
54. Drissi-Daoudi L, Doerig A, Herzog MH. Feature integration within discrete time windows. *Nat Commun*. 2019;10(1):1-8.
55. Toet A, Levi DM. The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Res*. 1992;32(7):1349-57.
56. 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.
57. 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;
58. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Springer; 2014. p. 818-33.
59. 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.
60. Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K. Detectron. 2018.
61. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol*. 2014;10(11):e1003915.
62. Kietzmann TC, McClure P, Kriegeskorte N. Deep neural networks in computational neuroscience. *BioRxiv*. 2018;133504.
63. VanRullen R. Perception science in the age of deep neural networks. *Front Psychol*. 2017;8:142.
64. 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.
65. 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.
66. Mathôt S, Schreij D, Theeuwes J. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behav Res Methods*. 2012;44(2):314-24.

General discussion

Summary of the results

For decades, human vision was studied mainly in terms of low-level circuits, using electrophysiological and neuro-anatomical measurements as constraints for successful models (1–4). This approach was fruitful, as it allowed to discover important fundamental mechanisms and paved the road for future vision research. Based on these discoveries, human vision was modelled as a feedforward hierarchy of increasingly complex representations (5). The models built upon this framework were able to explain how humans can recognize complex objects with low computational costs and few resources (5–12). These efforts resulted in the recent development of deep convolutional neural networks, state-of-the-art models both in computer vision (13–17) and neuroscience (18–25).

However, relying only on feedforward models and low-level circuits measurements misses important aspects of human vision since, as shown, these circuits cannot be studied in isolation (26). One reason is that recurrent connections play a central role in the visual cortex. For example, feedback occurs at the very first stage of vision. Indeed, V1 neurons respond to illusory contours only after they are formed in V2 (27–29). Moreover, neuron tuning properties are considerably altered by the presence of contextual elements, as exposed in V1 (30,31) and V4 (32).

Although many object recognition tasks can be solved by feedforward circuits using very simple computations, this is only a small subset of what vision does. Human vision has evolved to encapsulate many intricate behavioural functions. It is thus likely subject to more constraints than object recognition performance only, motivating the use of recurrent connections. In particular, recurrent connections in the visual cortex might allow the integration of contextual information in the representation of natural visual stimuli (33). Although every recurrent network has an equivalent unfolded feedforward network, the former use orders of magnitude fewer parameters and neurons (34). Hence, given the importance of recurrent connections in the human visual cortex, vision cannot be reduced to basic computations. It is important to probe high-level properties of human vision using adapted paradigms.

Along these lines, visual crowding can be used as a probe into human vision. Based on the success of feedforward models, it is usually described as the consequence of mandatory pooling that occurs along the processing hierarchy necessary to perform object recognition

(35–40). This explanation was validated by comparing the predictions of feedforward pooling models to different hallmarks of visual crowding, such as Bouma’s law (41).

However, crowding paradigms involving more complex displays undermined the success of feedforward models of crowding. First, some hallmarks have been found to no longer be true. For example, Manassi et al. (42–45) showed that crowding in a vernier discrimination task can be strongly affected by elements that lie far beyond Bouma’s window, and that adding more flankers can release the target from crowding (uncrowding). Importantly, high-level information about the global layout of the flankers affects low-level information. Configuration changes occurring in a range of almost 20 degrees determine the perception of a vernier target that depends on few arcmins offsets (45). Second, hallmarks of crowding can be explained as the consequence of the limited resolution of selective visual attention (46–48). Selective attention can arise both from top-down or bottom-up factors (49). Given the relative balance between feedforward and recurrent connections in the visual cortex, crowding and its hallmarks may as well be the consequence of, or at least heavily influenced by top-down processes.

These results raise the possibility that feedforward models of vision lack global computations and top-down processing. The results in this thesis show that local feedforward models cannot reproduce global effects in crowding. In contrast, it is shown that only the models that include grouping and segmentation processes reproduce human behaviour. Taken together, these results highlight the importance of visual grouping in models of human vision.

In Chapter 1, we used a large battery of stimuli in which the global flanker layout has a strong impact on crowding strength as measured in behavioural experiments. We tested whether different models could predict the amount of crowding associated to different flanker configurations. More specifically, in half of the stimuli, flankers far outside Bouma’s window release the target from crowding in human data (uncrowding).

The results showed that the most informative characteristic in predicting the success of a model was whether the model contains a grouping stage or not. More precisely, amongst all tested models, the only model that reproduced human behaviour consistently (i.e., not due to overfitting) contained a dedicated grouping stage. Different models that include recurrent connections but do not implement grouping did not reproduce human behaviour. This

suggests that grouping and segmentation processes are needed to explain global properties of visual crowding and human vision in general.

In the “winning” model (Laminart model; 50), segmentation is initiated locally by a top-down signal that triggers the networks dynamics before segmentation spreads along connected contours. This top-down signal can be interpreted as the tendency of human observers to segment the visual stimulus as efficiently as possible to perform the vernier discrimination task. It could be argued that this is in contradiction with a large body of literature showing that perceptual grouping may occur independently of top-down attentional selection (51–54). However, selection signals in the Laminart model need not be top-down. In a recent conference proceeding, it was shown that uncrowding could as well arise from bottom-up salience in the Laminart model (55).

Other models of grouping capture human behaviour as well, as shown by a different study that measured the performance of capsule networks. A capsule network is a type of deep convolutional network adding recurrent processing to implement grouping and segmentation, using the same stimuli (56). Importantly, the authors compared the behaviour of capsule networks to different types of deep neural networks using the same number of parameters, but that did not implement grouping processes (one purely feedforward, one with lateral connections and one with recurrent connections). Uncrowding occurred only for capsule networks, suggesting that recurrent processing itself is not sufficient to reproduce the effects of configuration in crowding. Importantly, uncrowding was effective only after a certain number of recurrent iterative loops, meaning that recurrent processes are necessary to reproduce global configuration effects in crowding.

In Chapter 2, we tested whether the Texture Tiling model (57) captures global configuration effects in crowding. Rosenholtz et al. (57) argued that visual grouping and segmentation processes are not needed to reproduce these effects. They showed that information about the spatial configuration of flankers passes through the high-dimensional pooling stage of their model. This information can later yield uncrowding effects at the decision process level, for example by reducing the target position uncertainty. This scenario does not require top-down computations (58).

The results of Chapter 2 showed that the texture tiling model fails to reproduce global configuration effects in crowding. Overall, the human performance in the original crowding experiments had no correlation with the performance of the model. Moreover, we showed that the behaviour of the model is equivalent to a simple pooling model. Indeed, in the Texture Tiling model, crowding increases in a monotone fashion, depending on the density of pixels in the flanking patterns. This is in contradiction with human results, in which flanker density is a weak predictor.

Along the same line, in a recent study, we dissected global effects in visual crowding by breaking down the flanker configurations that involve uncrowding (59). Vernier discrimination performance was measured in humans for different partitions of flanker configurations (for example, seven square flankers are split into two distinct configurations containing only the vertical or horizontal lines of the squares). First, crowding could not be explained by the performance of its parts (summing the effect of the flanker parts did not equal the effect of the whole), arguing against a pooling explanation. Second, three different models processing the input globally were validated on the data. One model was the Texture Tiling model, based on high-dimensional, low-level pooling (57). Both other models included grouping processes, either based on low-level feature integration (the Laminart model; 54) or on shape-level feedback (capsule network; 50,55). The texture tiling model reproduced none of the results, and the Laminart model only accounted for a subset of the data. The capsule network reproduced the whole set of human results, suggesting that object-like grouping is necessary to explain global effects in crowding.

In Chapter 3, emphasis was put on deep convolutional neural networks. The reason is that these networks are the most successful models in terms of behavioural performance in various vision-based tasks, in which they often outperform humans. In addition, their neurons' activity shows correlation with the neurons of the primate ventral stream (61–63), making them a potential model of human vision (18,24,25). In this study, we used different versions of deep convolutional neural networks pretrained on ImageNet (64), and trained simple classifiers to perform a vernier discrimination task to investigate whether these networks reproduce global effects observed in human crowding experiments (in this case, uncrowding).

We used AlexNet (15), because it is often compared to the human visual system (18,23) and ResNet-50 (65), a more sophisticated neural network that reaches better standards in object recognition and correlates better with primate cortical activity (66). Moreover, to investigate whether this failure was due to the general architecture of deep convolutional networks or to their training regime, we used a version of ResNet-50 trained on a modified version of ImageNet that forces the network to focus on global aspects of the stimulus when performing image recognition (67).

As a result, all tested networks showed crowding, but none reproduced uncrowding effects when adding more flankers. Occluding specific regions of the stimuli revealed that all networks focused on local features and ignored the overall configuration, consistent with a local pooling account of crowding in deep networks. Overall, the results of this study suggest that deep neural networks do not reproduce global aspects of visual crowding because they are based on local pooling, and not because of their training procedure.

Although both humans and deep neural networks are subject to visual crowding when performing object recognition tasks, the reasons for this impairment are fundamentally different. A consequence of this finding is that, although deep neural networks reach the same performance as humans in a large number of vision-based tasks, it cannot be concluded that the architecture and computations used to perform the tasks are the same. These results cast doubt on the proposed similarity between the neural activity of deep neural network and neurons in the primate ventral stream (61–63).

Along these lines, it was shown that a network trained on ImageNet and a randomly initialized counterpart show similar correlations to neurons in the mouse visual cortex (68). A more extensive study toned this claim down, but acknowledged that using deep neural networks as models of human vision was not as straightforward as it seems (69). As a side comment, it is still unknown what neural code is used by the visual cortex and whether the correlated activities are of relevance at all.

Recently, Lonqvist et al. (70), in a similar study as the one of Chapter 3, used crowding as a probe to compare the behaviour of deep convolutional neural networks and humans. Instead of focusing on effects of configuration, they manipulated parameters in simple crowding stimuli, such as size, target-flanker spacing and feature similarity. They also found that although

both humans and deep networks are subject to crowding, the underlying reasons for this breakdown are different. More specifically, they found that local pooling is the primary source of crowding in deep networks, independently of their specific architecture. They proposed that crowding in humans is unlike that in convolutional neural networks because there are many recurrent connections in the human visual cortex, which are absent from deep networks. The results in Chapter 3 add to this affirmation that only adding recurrent connections or global computations to traditional models of vision is not fully sufficient to reproduce all aspects of human-like object recognition. Indeed, it has been argued that ResNet-50 can be seen as an unfolded recurrent network (34). Our results show that this network does not reproduce uncrowding, and that training it to focus on global aspects of the visual input does not suffice either.

Inward-outward anisotropy is seen as a litmus-test for crowding (71). In the study of Lonqvist et al. (70), to enforce a fairer comparison between humans and networks, cortical magnification was added by deteriorating the resolution of the image input in the periphery. An inward-outward anisotropy was observed. However, the direction of the effect was the opposite of what is observed in humans. This suggests that simple pooling combined with cortical magnification does not account for the inward-outward anisotropy in visual crowding. Alternatively, cortical magnification was implemented in the population coding model of Van den Berg et al. (40) by scaling the neural population pooling range with eccentricity. In their model, a peripheral flanker induced more crowding than a foveal flanker, as in humans. Moreover, their model predicted the anisotropy to be strongest at intermediate target-flanker spacings, as observed in Farzin et al. (72). As another alternative, we modelled inward-outward anisotropy as the consequence of visual segmentation, combined to cortical magnification (73). We found that segmenting the target from the flankers is harder for a peripheral flanker because of its impoverished representation, producing the expected effect. Our model also predicted the anisotropy effect to be strongest at intermediate target-flanker spacing.

Here, two models based on pooling make opposite predictions for a specific hallmark of crowding, whereas two different accounts of crowding make a similar prediction. As a possible limitation of the latter model validation, the tested stimuli might only represent a tiny subset of the possible relevant configurations to measure. To eliminate this possible confound, it is

important to complete these validations using different paradigms in which the stimuli are determined in a bottom-up manner and in the future.

This is what was done in Chapter 4. This study used human data from Van der Burg et al. (74) in which visual crowding was measured in dense displays. Instead of sticking to specific hallmarks or paradigms, the stimuli were selected using a genetic algorithm (75). Importantly, the stimulus selection was made in a bottom-up manner. In the previous chapters, we used paradigms tailored to probe the global aspects of crowding. Here, we tested whether our model results also hold true for stimuli that were not designed to highlight the importance of grouping processes. Importantly, the human data of Van der Burg (74) suggest that Bouma's window shrinks to the nearest neighbour distance, which is the exact opposite effect compared to the human data used in Chapter 1 to 3, in which elements beyond Bouma's window affect crowding strength.

The results of Chapter 4 are in line with the previous chapters. All models based on pooling cannot reproduce the human data, whereas all models that include grouping and segmentation processes can explain the data. Putting together the latter results with the ones in Chapter 1, visual grouping better explains the range of interaction between visual elements than Bouma's window. Crucially, a two-staged model in which a segmentation model (Laminart model; 54) is connected to the population coding model of Van den Berg et al. (40), based on feedforward pooling, led to the best results and proved that a happy marriage can be sealed between both classes of models.

It is possible to model the human visual system in many ways. In this thesis, the main approach that we chose was to pit different classes of models against each other, based on the same sets of stimuli and paradigms. This approach can be related to comparative biology, in which small genotypic differences lead to drastic phenotype differences. Species-fair comparisons are the key to avoid drawing misrepresentative conclusions using this approach (76). Here, a first step towards fair model comparisons was to select paradigms and stimuli in both top-down and bottom-up fashions.

Limitations

In this thesis, it was shown that models need to include a grouping stage to explain global aspects of crowding. However, although model results are unequivocal, grouping strength between target and flankers is hard to assess quantitatively in a human experiment. Hence, a direct correlation between grouping strength and uncrowding is complex to establish. To answer this concern, subjective ratings can be used to assess how much the target stands out from different flanker configurations in humans. However, this measure may be subject to biases from different strategies that observers use to estimate grouping strength.

Another limitation is that the results in this thesis do not point towards specific mechanisms that would explain *how* the global configuration affects target visual acuity in crowding paradigms and it is yet unclear how grouping models would account for the classic effects of crowding. In particular, do far away objects only ever improve performance, or can they interfere and deteriorate performance further, even when beyond the known interference range? In the former case, the classic interference mechanisms would operate only after global spatial processing, and pooling would occur only within groups. In the latter case, interference would occur during global spatial processing, by an unknown and more complex mechanism that would require further research. Evidence from “super-crowding” paradigms (77–79) speaks towards the latter case, since elements beyond Bouma’s window can further deteriorate performance. However, these effects are restricted to a smaller range than what is attributed to uncrowding effects. Since the question is still pending, it is important to propose and compare different mechanisms for how elements may be grouped during neural processing. The networks that were used as representative of grouping models in the current thesis (Laminart model, Capsule network) come with their own weaknesses. Artificial vision research has come with many different architectures dedicated to performing image segmentation, which were not tested in the frame of this thesis (80–87).

Finally, although the modelled paradigms involve feature integration across large portions of space, they do not involve feature integration along the dimension of time. Using new paradigms in which time-based integration is observed in humans would allow to compare predictions from different hypotheses and mechanisms as models of visual grouping. An appealing example of such a paradigm is given in the next section.

Prospects

To come back to Marr's tri-level framework developed in the Introduction (88,89), the results in this thesis mainly addressed algorithmic questions about the visual system. Recurrent connections in models of vision should implement grouping and segmentation processes to reproduce human behaviour. Now the question arises of why such processes are crucial to human vision (Marr's computational level). In other words, if feedforward models of vision reach high levels of performance in many complex vision tasks, why should recurrences be of any need at all? There are two answers to this question. On the one hand, recurrences could only play a superficial role in vision, which may explain why neuroscience and artificial vision has quickly converged to purely feedforward networks. On the other hand, recurrences in the visual cortex might play a central role by allowing efficient computations in terms of space, time, and energy, as well as rich dynamics which can be used to generalize well to new tasks while relying on a small set of training data (90).

The results in this thesis support a view of the visual system in which the interplay between bottom-up processing and top-down inferences is not restricted to simple modulative effects, but rather a central mechanism of perception, in line with Bayesian hierarchical models of vision (91–96). Recently, CNNs with added feedback attentional processes were proposed, in which high-level context-dependent information drives lower layers' activations (97,98). In these models, the feedforward sweep lets all information flow to the upper layers, while feedback loops close certain gates according to higher-level activation, as a top-down feature-selective salience map. This mechanism reinforces relevant information and leads to better object recognition and localization performance in cluttered environments. This can be linked directly to visual crowding, in which performance is impaired in cluttered environment, but can be rescued by higher-level information.

There are various models that endorse this framework. The results in the current thesis suggest that the interplay between the top-down and bottom-up tracks should implement grouping and segmentation processes to reproduce human behaviour. However, as developed in the previous section, these results do not point towards a particular implementation or connectivity in the visual cortex. What exactly is defined by visual grouping in this context and what are the mechanisms that support human-like behaviour? Does it occur as the

consequence of object-level top-down activity, or from time-consuming local recurrent computations?

To tackle these questions, more advanced psychophysical paradigms may be needed. In the current thesis, it was shown that focusing on simple experimental paradigms to study human vision might be incomplete, because this does not reflect the complex ontology of its outputs and functions. Human vision has evolved in complex environments, in which elements are almost never presented in isolation. To capture the nature of the functions and connectivity it has developed, it is essential to use paradigms that reflect the natural settings of human vision. In Chapters 1 to 4, the importance of visual grouping was studied by using paradigms that span large portions of the visual field. However, although these paradigms highlight feature integration across space, they ignore that vision exists in time. They cannot investigate, for example, the importance of visual grouping across different instants in time. Well-controlled psychophysical paradigms in which elements span large portions of space and time might help to further characterize the grouping mechanisms that occur in human vision.

A good example is the sequential meta-contrast paradigm (SQM; 80,81). In this paradigm, observers are shown a sequence of frames consisting of two diverging streams of lines, starting with a single central line (Fig 1c). When the central line is offset, observers perceive the offset in the entirety of the stream to which they attend. Drissi-Daoudi et al. (101) investigated feature integration across time by including a second offset line later in the stream. If both offsets are opposed, the stream is perceived without offset. If they are the same, the stream is perceived with a larger offset. Importantly, this feature integration occurs until the offset frames are separated by up to 450 milliseconds. Moreover, when a third offset is included in a frame that is just outside this time-window, it is not integrated in the stream. It is as if vernier-offset features are mandatorily integrated across space and time by continuous subconscious processing, and that the brain “waits” for nearly half a second before it reaches a discrete conscious percept afterwards.

Low-level bottom-up processes cannot explain these results because the periods involved are way longer than any known bottom-up integration mechanism, such as visual persistence (102). Moreover, attention determines whether and how low-level information is integrated. Drissi-Daoudi et al. (101) proposed that in this paradigm the spatio-temporal continuity

between the line elements creates a grouped percept that is stabilized thanks to higher-level processes. Integration across large periods of time would be used by the brain, for example, to perform object segmentation over a cluttered background or behind intermittent occlusions.

Modelling integration over large periods of time and using dynamical stimuli requires good assumptions about the real-time flow of information in the visual cortex. A first step towards this goal is to take neuron transmission delays in hierarchical and recurrent models of vision. In such models, setting the physical duration of one input frame following real physiological axonal delays and brain oscillatory dynamics as measured in the visual cortex can link network computations to the real timing involved in dynamical stimuli (103).

Predictive coding, a type of multi-layered and recurrent model, can be seen as an implementation of the hierarchical Bayesian framework described above (104). This type of model provides promising insights regarding the results in the SQM paradigm. In predictive coding, bottom-up input is minimized through inhibition by top-down expectations and acts as an unsupervised prediction error signal (105). Lotter et al. (106) proposed a deep predictive coding network in which the role of top-down activity is to predict future input frames in realistic video streams (Fig 1a). Since objects are usually composed of elements that move together in natural video streams, this network can be seen as a model of object-level grouping.

Recently, Hogendoom and Burkitt (107) proposed to incorporate axonal delays in the predictive coding framework. They suggest that minimizing prediction error and predicting future frames in such a network might cause both bottom-up and top-down connections to carry information about the location and speed of visual elements. This may facilitate the synchronization of representations throughout all layers of the hierarchy (Fig 1b), similarly to the emergence of the dual stream architecture in the primate visual cortex (108). Importantly, because of this synchronization process, bottom-up connections would carry extrapolations about the future states of the stimulus. In other words, even the very first frame following stimulus onset would never reach higher levels of the hierarchy in its intact form, forcing top-down processes to integrate low-level information over large periods of time.

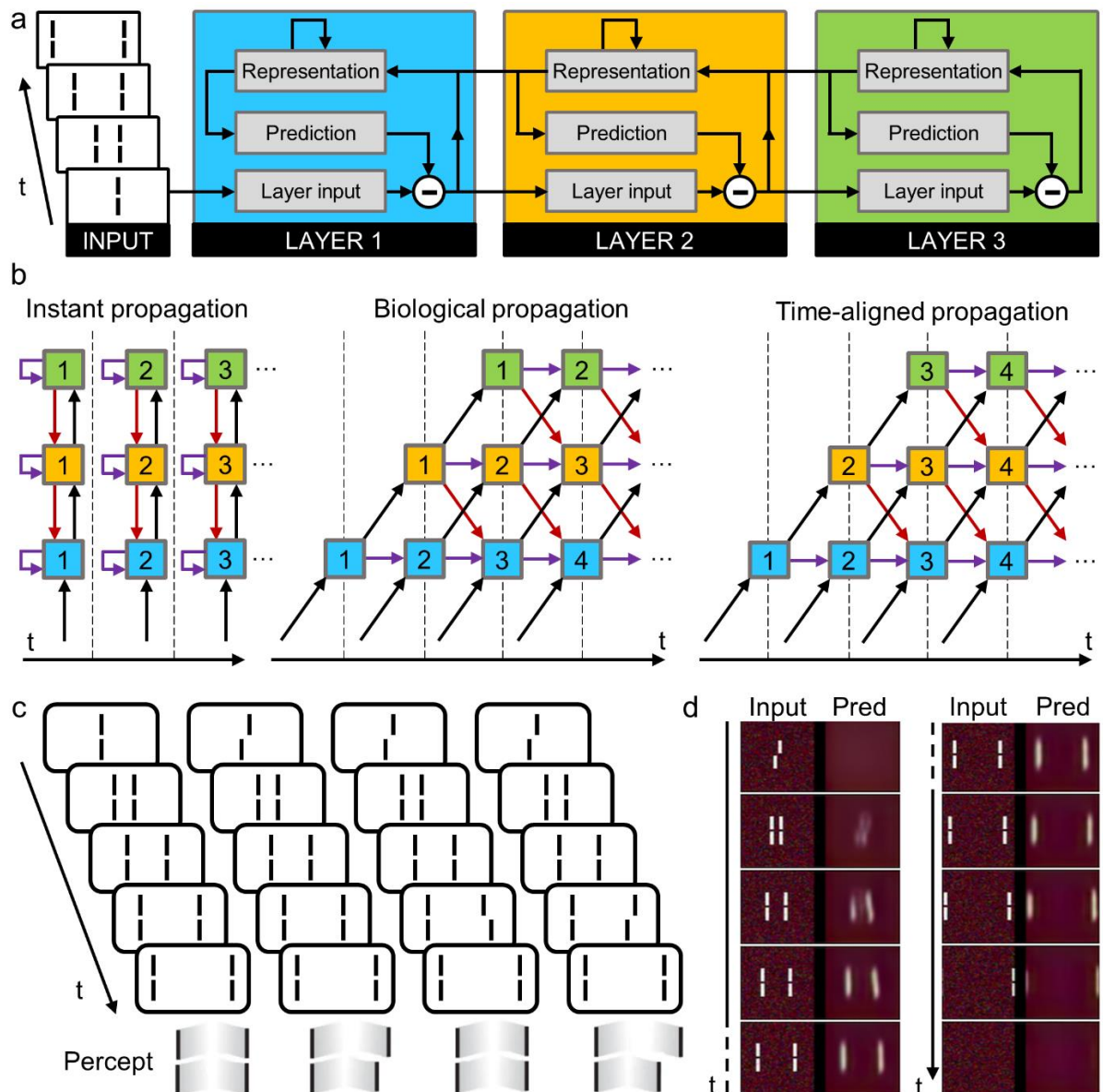


Fig 1. a. Deep predictive coding network for future frame predictions, as proposed by Lotter et al. (106). In the bottom-up pass, the output of each layer is the difference between the prediction of its own input and the actual received input. In the top-down pass, each layer updates its own representation states, based on higher-level information, to generate latent space predictions. **b.** Different ways of updating the states and inputs of the network. The vertical dashed-lines separate the network time-step simulations. **Left.** Classic propagation of information, as done in Lotter et al. (106). At each time step, visual information propagates through the whole network (full bottom-up and top-down pass). **Center.** The propagation of information takes axonal delays into account. Every (bottom-up, top-down or lateral) connection takes one time-step to propagate information. **Right.** After training, if prediction error is minimal, information is synchronized throughout the layers. This implies that top-down and bottom-up connections carry extrapolations about future visual states. **c.** SQM paradigm. Different versions give rise to different conscious percepts. **d.** Input and output of the predictive coding network of Lotter et al. (106) trained to perform next frame prediction with instant propagation of information. Integration occurs in the network but does not last more than one frame (90; ongoing work).

Although long periods of integrations might be produced, this kind of network cannot explain the emergence of discrete percepts, such as the ones highlighted by the results of the SQM paradigm. For this, a threshold mechanism would be needed, such as in a drift diffusion model (110). Alternatively, attractor states could be formed, for example by forcing the network to converge to stable states during training, even when the input is highly dynamical (contractor recurrent back-propagation; 92). The discrete percepts would arise as the network jumps between stable states, driven by a dynamical input that overcomes its stability. The length of the window of integration would be set by the amount of stability that the network acquired during training.

To summarize, adding biological realism to hierarchical and recurrent models of the visual cortex will potentially provide a fertile ground for new ideas and facilitate investigation on how human-like visual grouping works in space and time, hopefully contributing to the conception of better models of human vision.

Conclusion

In this thesis, visual crowding was used as a testbed to probe human visual processing. It was shown that focusing only on feedforward models and low-level circuits' measurements potentially misses important aspects of human vision. Comparative modelling studies were performed and showed that mid-level grouping and segmentation processes are crucial to understand global effects in visual crowding. Importantly, these effects are not idiosyncratic. As shown in the Introduction, they rather reflect a general and ubiquitous strategy of perception to process information efficiently when large portions of the visual field are involved. Hence, the results in this thesis do not apply to (un)crowding paradigms only. They provide constraints for future artificial vision models. It becomes more and more clear that adding recurrences to artificial models of vision can improve their performance, as well as their biological plausibility. However, recurrent connections are extremely under-constrained, and feedforward models still reach higher levels of performance in many complex visual tasks. Here, it was shown that implementing grouping and segmentation processes using recurrent connections reproduces human behaviour in complex settings. Hence, this simple addition may help state-of-the-art networks to process large parts of the visual input more efficiently and more robustly.

This work can be embedded in a more general effort to understand the complexity of the brain. Modelling the brain boils down to identifying the inductive biases that evolution has found to reach sufficient performance on a large variety of complex tasks, relying on the most efficient computations that comply with the physical constraints of perceptual systems. It has been proposed that, to expose these biases, vision research should focus on three essential components of network systems: objective functions, learning rules and architecture (112).

One striking example of inductive bias that drove vision research forward is the translational invariance of semantic content in images, constraining artificial networks to implementing local feature detectors with convolutional filters. This reduced the networks' number of parameters needed to perform object recognition by several orders of magnitude and allowed convergence to very high levels of performance in a short amount of time. This led to the success of feedforward models of human vision.

Another example is the use of simple explanations to make sense of the outer world. Here, the constraint is to encode information as efficiently as possible, i.e., to minimize an objective

function defined by the amount of activity in the whole network as, for example, in predictive coding. Visual grouping is well embedded in this idea, since object-level feedback helps explain a lot of neural activity, especially with dynamic inputs.

Finding and searching for these inductive biases helps us understand how network computations may reflect the natural symmetries of the outer world, reminiscent of Noether's theorem, in which every symmetry of a physical system has a corresponding constraint, i.e., a conservation law (113,114). For example, in the case of the translational symmetry of objects in the visual field (an object's identity does not change if translated in the visual field), an associated constraint can be observed in perceptual systems trained on image recognition: the rapid emergence of using shared convolutional filters, as observed in the primate visual cortex as well as in the success of deep convolutional networks. Following this path may hopefully lead vision research to at least a fraction of the success that was achieved by physics in the last century.

References

1. DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? *Neuron*. 2012;73(3):415–34.
2. 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.
3. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*. 1962;160(1):106.
4. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. *J Physiol*. 1968;195(1):215–43.
5. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci*. 1999;2(11):1019–25.
6. VanRullen R. The power of the feed-forward sweep. *Adv Cogn Psychol*. 2007;3(1–2):167.
7. Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T. A quantitative theory of immediate visual recognition. *Prog Brain Res*. 2007;165:33–56.
8. Perrett DI, Oram MW. Neurophysiology of shape processing. *Image Vis Comput*. 1993;11(6):317–33.
9. Ullman S, Basri R. Recognition by linear combination of models. MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB; 1989.
10. Wallis G, Rolls ET. Invariant face and object recognition in the visual system. *Prog Neurobiol*. 1997;51(2):167–94.
11. Hummel JE, Stankiewicz BJ. An architecture for rapid, hierarchical structural description. *Atten Perform XVI Inf Integr Percept Commun*. 1996;93–121.
12. Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: *Competition and cooperation in neural nets*. Springer; 1982. p. 267–85.
13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: *Advances in neural information processing systems*. 2014. p. 2672–80.
14. Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K. Detectron. 2018.
15. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
16. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. *ArXiv Prepr ArXiv181204948*. 2018;
17. 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.
18. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol*. 2014;10(11):e1003915.
19. Nayebi A, Bear D, Kumbhani J, Kar K, Ganguli S, Sussillo D, et al. Task-Driven Convolutional Recurrent Models of the Visual System. *ArXiv Prepr ArXiv180700053*. 2018;
20. 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.
21. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 586–95.
22. 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;
23. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Springer; 2014. p. 818–33.
24. Kietzmann TC, McClure P, Kriegeskorte N. Deep neural networks in computational neuroscience. *BioRxiv*. 2018;133504.
25. VanRullen R. Perception science in the age of deep neural networks. *Front Psychol*. 2017;8:142.

26. Olshausen BA, Field DJ. What is the other 85 percent of V1 doing. Van Hemmen T Sejnowski Eds. 2006;23:182–211.
27. Roelfsema PR, Lamme VA, Spekreijse H. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*. 1998;395(6700):376–81.
28. Lee TS, Nguyen M. Dynamics of subjective contour formation in the early visual cortex. *Proc Natl Acad Sci*. 2001;98(4):1907–11.
29. Lee TS, Mumford D, Romero R, Lamme VA. The role of the primary visual cortex in higher level vision. *Vision Res*. 1998;38(15–16):2429–54.
30. David SV, Vinje WE, Gallant JL. Natural stimulus statistics alter the receptive field structure of v1 neurons. *J Neurosci*. 2004;24(31):6991–7006.
31. David SV, Gallant JL. Predicting neuronal responses during natural vision. *Netw Comput Neural Syst*. 2005;16(2–3):239–60.
32. Benjamin AS, Ramkumar P, Fernandes H, Smith MA, Kording KP. Hue tuning curves in V4 change with visual context. *bioRxiv*. 2019;780478.
33. Boutin V, Franciosini A, Chavane F, Ruffier F, Perrinet L. Sparse Deep Predictive Coding captures contour integration capabilities of the early visual system. *ArXiv Prepr ArXiv190207651*. 2019;
34. Liao Q, Poggio T. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *ArXiv Prepr ArXiv160403640*. 2016;
35. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis*. 2009;9(12):13–13.
36. 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.
37. Wilkinson F, Wilson HR, Ellemberg D. Lateral interactions in peripherally viewed texture arrays. *Josa A*. 1997;14(9):2057–68.
38. Greenwood JA, Bex PJ, Dakin SC. Positional averaging explains crowding with letter-like stimuli. *Proc Natl Acad Sci*. 2009;106(31):13130–5.
39. Nandy AS, Tjan BS. Saccade-confounded image statistics explain visual crowding. *Nat Neurosci*. 2012 Mar;15(3):463–9.
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.
41. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Res*. 1973;13(4):767–82.
42. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis*. 2012;12(10):13–13.
43. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. *J Vis*. 2013;13(13):10–10.
44. Manassi M, Hermens F, Francis G, Herzog MH. Release of crowding by pattern completion. *J Vis*. 2015;15(8):16–16.
45. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis*. 2016;16(3):35–35.
46. Intriligator J, Cavanagh P. The spatial resolution of visual attention. *Cognit Psychol*. 2001;43(3):171–216.
47. He S, Cavanagh P, Intriligator J. Attentional resolution and the locus of visual awareness. *Nature*. 1996;383(6598):334–7.
48. Chakravarthi R, Cavanagh P. Temporal properties of the polarity advantage effect in crowding. *J Vis*. 2007;7(2):11–11.
49. Beck DM, Kastner S. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Res*. 2009;49(10):1154–65.
50. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychol Rev*. 2017;124(4):483.

51. Carther-Krone TA, Lawrence-Dewar JM, Shomstein S, Nah JC, Collegio AJ, Marotta JJ. Neural Correlates of Perceptual Grouping Under Conditions of Inattention and Divided Attention. *Perception*. 2020;49(5):495–514.
52. Driver J, Davis G, Russell C, Turatto M, Freeman E. Segmentation, attention and phenomenal visual objects. *Cognition*. 2001;80(1–2):61–95.
53. Kimchi R, Peterson MA. Figure-ground segmentation can occur without attention. *Psychol Sci*. 2008;19(7):660–8.
54. Moore CM, Egeth H. Perception without attention: Evidence of grouping under conditions of inattention. *J Exp Psychol Hum Percept Perform*. 1997;23(2):339.
55. Bornet A, Kroner A, Kaiser J, Scholz F, Francis G, Herzog M. Using the Neurorobotics platform to explain global processing in visual crowding. In: 41st European Conference on Visual Perception (ECVP). 2018.
56. 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.
57. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. *J Vis*. 2019;19(7):15–15.
58. Reuther J, Chakravarthi R. Response selection modulates crowding: a cautionary tale for invoking top-down explanations. *Atten Percept Psychophys*. 2019;1–16.
59. Choung OH, Bornet A, Doerig A, Herzog MH. Dissecting (un)crowding. (submitted);
60. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: *Advances in neural information processing systems*. 2017. p. 3856–66.
61. Nayebi A, Bear D, Kubilius J, Kar K, Ganguli S, Sussillo D, et al. Task-Driven Convolutional Recurrent Models of the Visual System. *ArXiv Prepr ArXiv180700053*. 2018;
62. 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.
63. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018. p. 586–95.
64. 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.
65. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770–8.
66. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*. 2018;407007.
67. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Prepr ArXiv181112231*. 2018;
68. Cadena SA, Sinz FH, Muhammad T, Froudarakis E, Cobos E, Walker EY, et al. How well do deep neural networks trained on object recognition characterize the mouse visual system? 2019;
69. Conwell C, Alvarez G. Is Rodent Visual Cortex Really Just a Randomly Initialized Neural Network? *J Vis*. 2020;20(11):968–968.
70. Lonqvist B, Clarke AD, Chakravarthi R. Crowding in humans is unlike that in convolutional neural networks. *Neural Netw*. 2020;
71. Petrov Y, Popple AV, McKee SP. Crowding and surround suppression: Not to be confused. *J Vis*. 2007;7(2):12–12.
72. Farzin F, Rivera SM, Whitney D. Holistic crowding of Mooney faces. *J Vis*. 2009;9(6):18–18.
73. Bornet A, Kaiser J, Kroner A, Falotico E, Ambrosano A, Cantero K, et al. Running large-scale simulations on the Neurorobotics Platform to understand vision-the case of visual crowding. *Front Neurorobotics*. 2019;13:33.
74. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. *J Exp Psychol Hum Percept Perform*. 2017;43(4):690.

75. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press; 1992.
76. Firestone C. Performance vs. competence in human–machine comparisons. *Proc Natl Acad Sci*. 2020;117(43):26562–71.
77. Rashal E, Yeshurun Y. Contrast dissimilarity effects on crowding are not simply another case of target saliency. *J Vis*. 2014;14(6):9–9.
78. Soo L, Chakravarthi R, Andersen SK. Critical resolution: A superior measure of crowding. *Vision Res*. 2018;153:13–23.
79. Vickery TJ, Shim WM, Chakravarthi R, Jiang YV, Luedeman R. Supercrowding: Weakly masking a target expands the range of crowding. *J Vis*. 2009;9(2):12–12.
80. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on u-net (R2U-net) for medical image segmentation. *ArXiv Prepr ArXiv180206955*. 2018;
81. Linsley D, Kim J, Ashok A, Serre T. Recurrent neural circuits for contour detection. *ArXiv Prepr ArXiv201015314*. 2020;
82. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *ArXiv Prepr ArXiv170605587*. 2017;
83. Badrinarayanan V, Handa A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *ArXiv Prepr ArXiv150507293*. 2015;
84. Liu W, Rabinovich A, Berg AC. Parsenet: Looking wider to see better. *ArXiv Prepr ArXiv150604579*. 2015;
85. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. *ArXiv Prepr ArXiv151107122*. 2015;
86. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 2881–90.
87. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE international conference on computer vision*. 2015. p. 1520–8.
88. Marr D. *Vision: A computational investigation into the human representation and processing of visual information*. 1982;
89. McClamrock R. Marr’s three levels: A re-evaluation. *Minds Mach*. 1991;1(2):185–96.
90. van Bergen RS, Kriegeskorte N. Going in circles is the way forward: the role of recurrence in visual inference. *Curr Opin Neurobiol*. 2020;65:176–93.
91. Dayan P, Hinton GE, Neal RM, Zemel RS. The helmholtz machine. *Neural Comput*. 1995;7(5):889–904.
92. Mumford D. On the computational architecture of the neocortex. *Biol Cybern*. 1992;66(3):241–51.
93. Knill DC, Richards W. *Perception as Bayesian inference*. Cambridge University Press; 1996.
94. Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. *Annu Rev Psychol*. 2004;55:271–304.
95. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. *JOSA A*. 2003;20(7):1434–48.
96. George D, Hawkins J. A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In: *Proceedings 2005 IEEE International Joint Conference on Neural Networks, 2005. IEEE; 2005*. p. 1812–7.
97. Chikkerur S, Serre T, Tan C, Poggio T. What and where: A Bayesian inference theory of attention. *Vision Res*. 2010;50(22):2233–47.
98. Cao C, Liu X, Yang Y, Yu Y, Wang J, Wang Z, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: *Proceedings of the IEEE international conference on computer vision*. 2015. p. 2956–64.
99. Otto TU, Ögmen H, Herzog MH. The flight path of the phoenix—The visible trace of invisible elements in human vision. *J Vis*. 2006 Aug 1;6(10):7–7.
100. Otto TU, Ögmen H, Herzog MH. Feature integration across space, time, and orientation. *J Exp Psychol Hum Percept Perform*. 2009;35(6):1670–86.
101. Drissi-Daoudi L, Doerig A, Herzog MH. Feature integration within discrete time windows. *Nat Commun*. 2019;10(1):1–8.

102. Enns JT, Lleras A, Moore CM. Object updating: A force for perceptual continuity and scene stability in human vision. *Space Time Percept Action*. 2010;503–20.
103. Alamia A, VanRullen R. Alpha oscillations and traveling waves: Signatures of predictive coding? *PLoS Biol*. 2019;17(10):e3000487.
104. Aitchison L, Lengyel M. With or without you: predictive coding and Bayesian inference in the brain. *Curr Opin Neurobiol*. 2017;46:219–27.
105. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci*. 1999;2(1):79–87.
106. Lotter W, Kreiman G, Cox D. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. *ArXiv160508104 Cs Q-Bio* [Internet]. 2017 Feb 28 [cited 2020 Jul 22]; Available from: <http://arxiv.org/abs/1605.08104>
107. Hogendoorn H, Burkitt AN. Predictive coding with neural transmission delays: a real-time temporal alignment hypothesis. *Eneuro*. 2019;6(2).
108. Ungerleider LG, Haxby JV. ‘What’ and ‘where’ in the human brain. *Curr Opin Neurobiol*. 1994;4(2):157–65.
109. Anselmet M. Modelling the Sequential Metacontrast Paradigm with Recurrent Neural Networks. 2020;
110. Ratcliff R. A theory of memory retrieval. *Psychol Rev*. 1978;85(2):59.
111. Linsley D, Ashok AK, Govindarajan LN, Liu R, Serre T. Stable and expressive recurrent vision models. *ArXiv Prepr ArXiv200511362*. 2020;
112. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. *Nat Neurosci*. 2019;22(11):1761–70.
113. Noether E. Nachrichten der Koniglichen Gesellschaft der Wissenschaften, Gottingen, Mathematisch-Physikalische Klasse 2, 235–257. *Invariante Var*. 1918;
114. Noether E. Invariant variation problems. *Transp Theory Stat Phys*. 1971;1(3):186–207.

Supplementary information

A: Supplementary information for Chapter 1

Back to Chapter 1: [\[Introduction\]](#) - [\[Methods\]](#) - [\[Results\]](#) - [\[Discussion\]](#)

SA1: Epitome model

Spatial extent	Mechanism	Organisation	Grouping component
Local	Substitution	Feedforward	No

In the Epitome model, described by Jojic et al. (1), large repeating patterns are summarized by small repeated representative image patches. Repeated patterns are substituted with their exemplars. The original image can subsequently be retrieved with good accuracy from the compressed representation, even though neighboring features encoded in the same patch are mingled. Epitomes are effectively a “*substitution*” model that exploits regularities. Although this model was not proposed as a model of crowding, it embodies many of the key characteristics of local pooling and substitution models.

Using the authors’ code available online (<http://www.cis.upenn.edu/~jshi/software/>) with the original parameters (designed to optimize image reconstruction accuracy for natural images and texture overlays), we ran the model on all stimuli. To evaluate performance, we (the authors) used the *classic texture* evaluation method, analysing the results qualitatively (see methods). In addition, we computed the model threshold as:

$$\iint_{x,y} |leftStim(x,y)| - |rightStim(x,y)| \, dx dy$$

where $leftStim(x,y)$ is the normalized intensity of pixel (x,y) in the left vernier offset version of the output. Effectively, this equation quantifies how different the normalized output images are for the left and the right vernier offset versions of the stimulus. If they are very different, the task is easy. Consistently across the dataset, the model successfully produces crowding but not uncrowding: performance was always worse when adding more flankers (Fig SA1). We suggest that the model cannot explain uncrowding because it compresses information from *local* regions of the image, ignoring global structure.

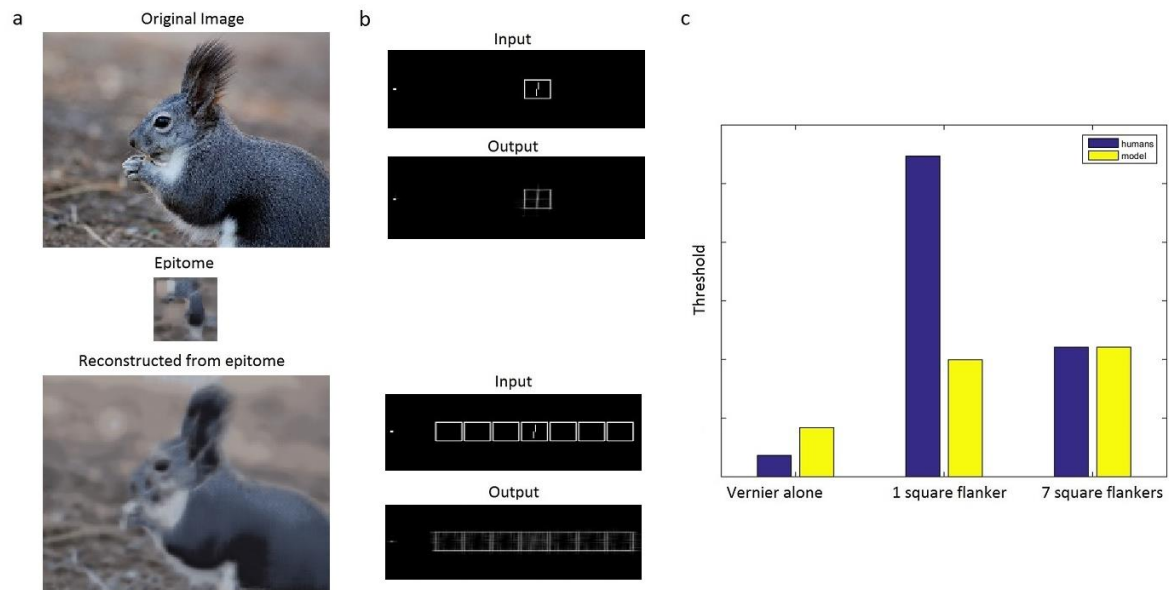


Fig SA1. Epitomes. **a.** Illustration of the epitome model. An image (left) is compressed into an epitome (center), a summary of local features. The image on the right is reconstructed from the epitome. **b.** As an example for the classic texture evaluation, we show the stimulus and reconstructed image for the 1- and 7-square conditions. Human vernier offset thresholds are better for the 1-square than the 7-square condition. The model does not produce uncrowding because vernier offset direction in the output is not easier to make out in the 7-square than in the 1-square case (according to the authors' judgment). **c.** Example for our performance measure. Human and model thresholds (see main text for how model threshold was computed) for vernier alone (condition 1), single square (condition 2) and 7 squares (condition 3). The 7-square threshold is higher than the 1- square threshold, in contrast with human performance. Note: the model outputs a number quantifying how different the left and right vernier offset versions of the input are (so the higher this difference, the better the performance). To make comparison with the human threshold easier, we applied the following monotonic transformation to the output: "threshold-like output" = $1/\text{"raw output"}$. Then, we scaled the result to be in the same range as the human results. This monotonic re-scaling cannot not change the conclusions because monotonic outputs are mapped on monotonic performance and the same is true for U-shaped functions (see methods).

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA2: Single texture model

Spatial extent	Mechanism	Organisation	Grouping component
Global	Pooling	Feedforward	No

Portilla & Simoncelli (2) proposed a set of statistics capable of capturing the key aspects of texture appearance to human vision (Fig SA3a). Balas et al., (3) suggested an explanation of crowding in which peripheral vision might measure these texture statistics in pooling regions

that overlap and tile the visual field. The intuition is that summary statistics provide an efficient way of extracting relevant information at low computational cost from natural images. Though Balas et al. proposed a model covering the entire visual field as described in the next subsection, they initially tested the predictions of a single pooling region, since texture synthesis procedures did not exist for multiple overlapping pooling regions. Each of their stimuli fell within a single Bouma-sized patch. They have since suggested that using a single pooling region, which greatly reduces computation time, can often suffice for texture-like stimuli that fall within a single pooling region (4).

Although the model was intended by Balas et al. to be applied only over a Bouma’s window-sized patch, we applied it to the entire stimulus to see if this kind of texture synthesis could capture long-range interactions between the vernier and other elements. The texture statistics are computed from pixel intensities taken from the entire image. Using the code provided by Portilla & Simoncelli (<https://github.com/LabForComputationalVision/textureSynth>), we created textures from all of our stimuli and the authors analyzed the results qualitatively using the *texture measure* (see Fig SA3c for two examples). The model produces strong crowding: vernier offsets are harder to discriminate from the textures when flankers are present. However, the model cannot explain uncrowding: consistently across our whole dataset, uncrowded conditions are worse than crowded conditions for this model (Fig SA3c). More elements always deteriorate performance. In their original contribution, Balas et al. seeded the texture synthesis algorithm using a low-pass, noisy version of the stimulus to reduce position noise. We also ran our stimuli using this method (see results repository online). While the output images became less distorted than without using the seed, it did not change the conclusion, because the target vernier remained much harder to detect in the textures synthesized from the uncrowded 7 flankers than from the crowded single flanker stimuli – i.e., there was no uncrowding.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA3: Texture tiling model (TTM)

Spatial extent	Mechanism	Organisation	Grouping component
Local	Pooling	Feedforward	No

The TTM model was first described by Balas et al. (3), with its first full instantiation developed by Freeman & Simoncelli (5). It computes summary statistics for overlapping local patches of the visual field, mimicking the way V2 receptive fields size grows with eccentricity (Fig SA3b). Balas, Rosenholtz and others have studied this model extensively, calling it the Texture Tiling Model (TTM; 31,32). In a series of papers, this model explained well the local aspects of visual tasks such as crowding and visual search. We ran a selection of stimuli through the TTM model (circles, squares, and irregular1). As with the previous textures, the results were analysed by the authors using the *classic texture measure*. Crowding was well captured, but uncrowding could not be explained by TTM (Fig SA3d). The vernier was not better represented as the number of flankers increased.

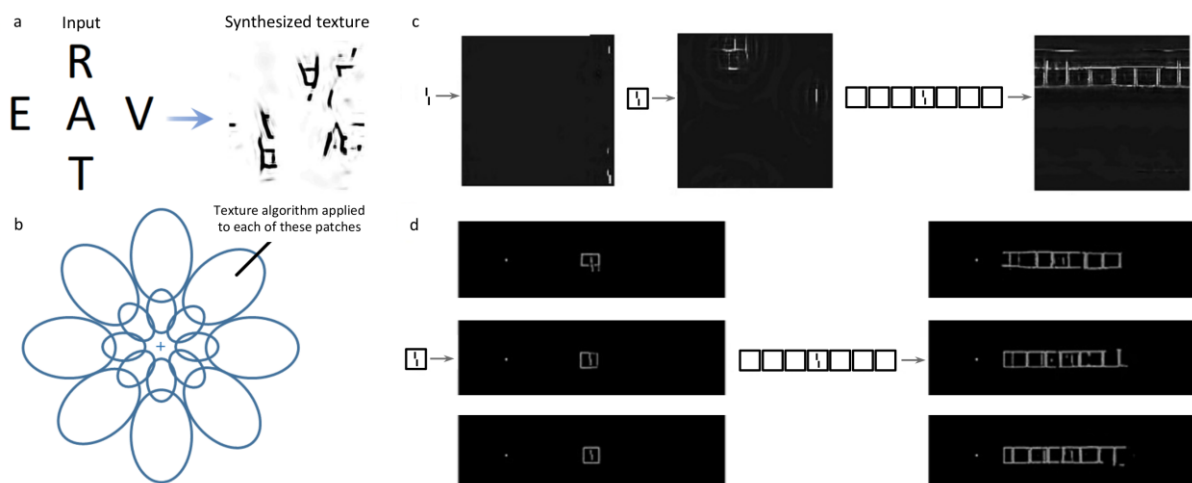


Fig SA3. Texture Synthesis and Texture Tiling Model. **a.** A texture (right) synthesized from the input on the left using the Portilla & Simoncelli (2) summary statistics. The output resembles crowding. Pooling- and substitution-like effects occur. **b.** Instead of applying the summary statistics process to the whole image at once, only local patches of the image are processed, yielding a local summary statistics model. The local patches are thought to reflect V2 receptive fields. **c.** Whole-field summary statistics. From left to right: stimuli and Portilla & Simoncelli textures for the vernier, 1-square and 7-square conditions. The vernier offset is easy to determine from the texture in the vernier alone condition, and slightly harder in the crowded condition (a right-offset is discernable in the middle top of the display). Across all data, the model consistently produces crowding, but no uncrowding, as exemplified in the right condition in which no offset is present at all. **d.** Texture Tiling model. The left column shows three synthesized examples from the 1-square condition. On the right is the 7-flanking squares case. The model cannot produce uncrowding: since the stimulus on the right is less crowded than on the left in the human data, the direction of the vernier should be easier to make out on the right than on the left. However, this is not the case.

We suggest that TTM alone cannot explain uncrowding because it is a sophisticated *local* mechanism that scrambles together neighboring elements. There is no mechanism allowing elements that do not share a pooling region with the target to directly affect the target representation. Our results suggest that neither pooling summary statistics over the entire stimulus nor pooling over previously tested local regions explains the behavioural results. If the whole field is used, uncrowding cannot occur because more elements mean more interference and thus worse performance. On the other hand, using local regions does not help because far away elements cannot improve performance in cases where humans show uncrowding.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA4: Deep textures

Spatial extent	Mechanism	Organisation	Grouping component
Global	Pooling	Feedforward	No

Gatys and colleagues (8) used deep neural networks to create textures. The algorithm starts with a noise image and iteratively modifies it to match the correlations between neuron activities in a set of layers. This procedure synthesizes textures that are often indistinguishable from the original image, creating true metamers (9). Deep textures were not intended to be applied to images like our stimuli, nevertheless we were interested in seeing if they could handle them because one could think of deep textures as synthesizing textures based on learned features rather than on the hand-coded features of Portilla & Simoncelli (2). Perhaps the learned features provide a better representation and thus better predict crowding.

Using Gatys et al.'s code (<https://github.com/leongatys/DeepTextures>) with their suggested set of parameters, we created textures of each stimulus in our database (Fig SA4a shows a selection of examples). We first evaluated model performance by the *classic texture measure* performed by the authors. Since the results were much less clear than for the previous texture approaches, we also conducted a psychophysical experiment with naive participants. Five subjects performed the *classic texture measure*: they were first explained the texture synthesizing process and then were shown textures synthesized from our stimuli. They were asked to report if they thought the texture was synthesized from a left- or right-vernier stimulus. We used three categories of stimuli (Gestalts, squares and circles), with ten textures

per stimulus (a total of 100 textures). Performance was at chance for all stimuli. Textures for the untested stimulus categories strongly resemble the tested categories (the vernier offset orientation is not visible in the textures, even for the vernier-alone condition). We tried different stimulus sizes, but this did not improve the results. In conclusion, despite its clear success at texture synthesis for natural images, the model in its present form is not suitable to study crowding with our stimuli.

Wallis et al. (10) have proposed a foveated model in which these deep statistics are computed over *local* image patches, just as the TTM computes Portilla and Simoncelli's statistics over local patches. The code is not yet publicly available, so we did not test it explicitly, however, we believe it will not explain uncrowding for exactly the same reasons that the TTM does not handle uncrowding better than Portilla and Simoncelli's whole field statistics: distant elements that are not in pooling regions around the target cannot affect the target representation.

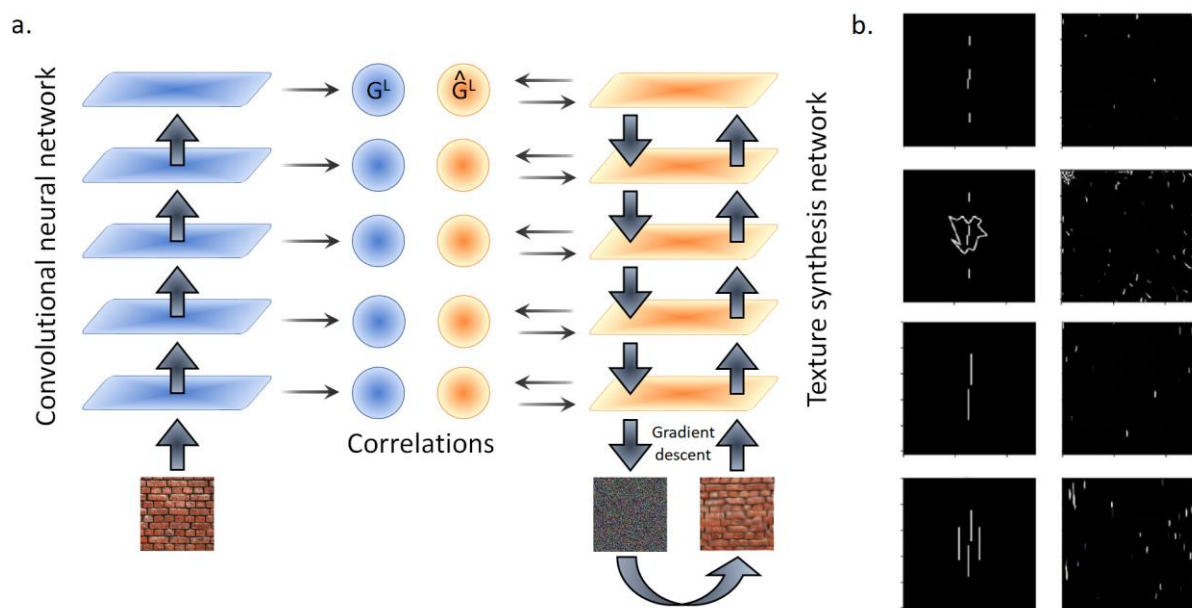


Fig SA4. Deep textures. **a.** In the deep texture algorithm, the correlation between a deep neural network's unit activities is used as a summary statistic. Textures are then synthesized to match that statistic. **b.** Original stimuli and textures synthesized from these stimuli using the deep textures algorithm by Gatys et al. (8). The vernier offset is poorly visible, therefore, despite its success at synthesizing textures, the model in its present form is not suitable to our stimuli. We tried different zooms on our stimuli, but the results did not change.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA5: Wilson & Cowan network with end-stopped receptive fields

Spatial extent	Mechanism	Organisation	Grouping component
Global	Pooling	Recurrent	No

Wilson & Cowan (11) proposed a mathematical model of simple cortical (excitatory and inhibitory) neurons interacting through recurrent lateral connexions. Variations of this kind of model have successfully accounted for visual masking data using stimuli similar to our *lines* category (12). We used a similar neural network for our crowding stimuli. The model first convolves the input image with an on-center, off-surround receptive field mimicking processing by the LGN. Next, the input activations are fed into both an excitatory and an inhibitory layer of neurons, which are reciprocally connected such that the excitatory units excite the inhibitory units and the inhibitory units inhibit the excitatory units. Details of the model, its filters, and its parameters can be found in (12) and (13). Although the filters are local, the strength of activity at any given pixel location partly depends on the global pattern of activity across the network because of the feedback connections. More generally, the feedback in the network functions like a discontinuity detector by enhancing discontinuities and suppressing regularities. Clarke, Herzog & Francis (14) applied this model to crowding stimuli, but it performed poorly and produced no uncrowding. For example, there was no difference between the stimuli in the Gestalts category and the length of the bars in the lines category had no effect at all on performance. Here, to improve the model, we replaced the classic receptive fields by end-stopped receptive fields so that each neuron is optimally activated only by stimuli of a specific length. There were three different sizes for the end-stopped receptive fields, corresponding to the size of a vernier bar, the size of the whole vernier, and the size of the flankers. To measure performance for each stimulus, for each end-stopped receptive field size, we took as output the state of the excitatory layer after stabilization (40 time-steps) and cross-correlated it with the vernier alone output. The cross-correlations for each end-stopped receptive field size were summed to yield a single output number per stimulus. We then fitted a psychometric function on one class of stimuli (training set) and used this function to provide model performance for all other classes of stimuli (testing set). Apart from the end-stopped receptive fields modification, we used the same parameters as in Hermens et al. (12).

We fit the psychometric function based on the model's output for the squares category, i.e., the squares category is the training set, and used this fit to measure performance on all other stimulus categories, i.e., all other categories are the testing set. We also tried to use each of the other categories as the training set; using the squares yielded the best results. The model produces crowding: performance drops in the presence of flankers. It also produces uncrowding but only for the training set (squares) and, to a lesser extent, for the irregular1 category. Indeed, performance is better in the 7 squares than in the single square condition (Fig SA5b), and marginally better in the 7 irregular1 than in the single irregular1 condition (Fig SA5c). For the other categories, there is no uncrowding (see Fig SA5d for an example). The choice of the training and testing sets has a strong influence on the conditions, which mimics human performance. Squares and lines are the categories for which size regularity seems to play the most important role. For all other classes, there is no uncrowding, regardless of the training set (circles, Gestalts, irregular2, hexagons, octagons, patternIrregular, patternStars & stars – Fig SA5c). This poor generalization capability suggests that the model uses idiosyncratic features of its training set rather than capturing general regularities, similar to overfitting.

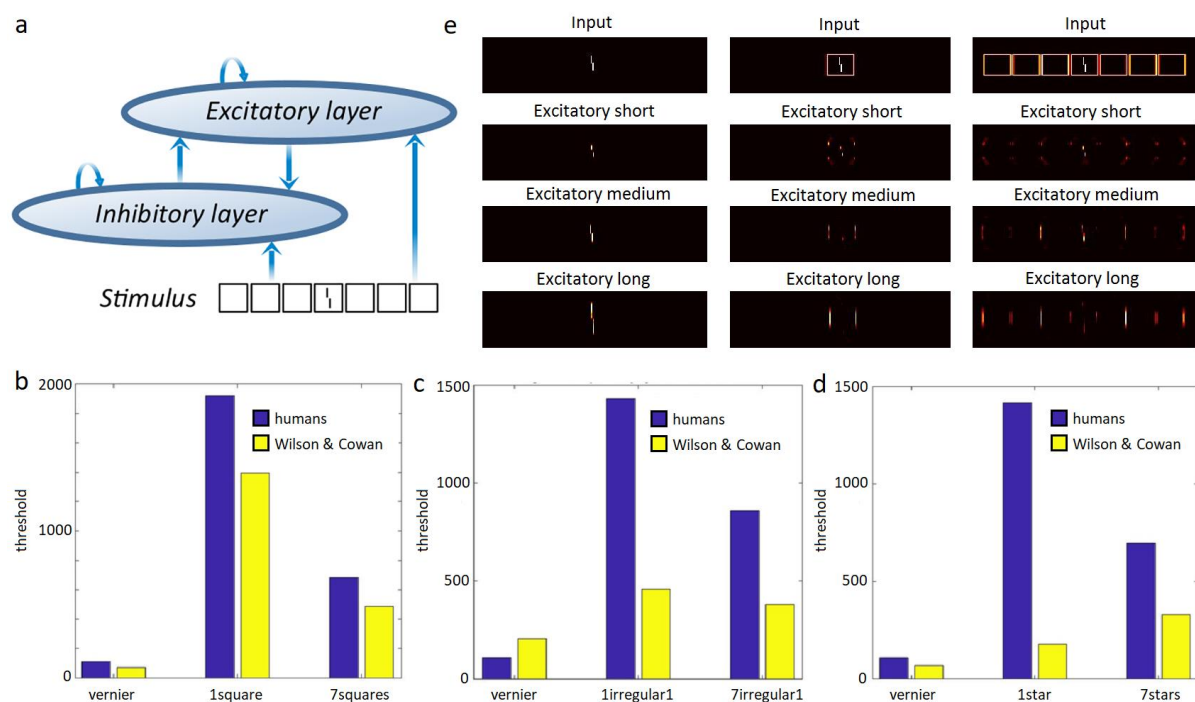


Fig SA5. Wilson and Cowan network with end-stopped receptive fields. **a.** Structure of the network in (12) which we augmented with end-stopped receptive fields. An excitatory and an inhibitory layer of neurons are activated by the stimulus and interact with one another. The output of the excitatory layer is cross correlated with a vernier template to measure performance. Figure from Hermens et al. (12). **b.** Output for the squares category (trained

on the squares category). In accordance with human results, performance is better in the 7 squares than in the 1 square case. **c.** Output for the irregular category (trained on the squares category). Performance is marginally better in the 7 irregular1 than in the 1 irregular1 case. **d.** Output for the stars category (trained on the squares category). There is no uncrowding for this stimulus. Uncrowding occurs only for specific kinds of stimuli, where element size regularities seem important. Further, performance depends strongly on which data are used for the training set, suggestive of overfitting. **e.** Model output images. Columns are different stimuli: vernier, 1 square and 7 squares. The first row shows the stimuli, and the three subsequent rows show the model output for the short, medium and long end-stopped receptive fields. The crucial result is that the vernier is *better* represented in the short and medium populations in the 7 squares than in the 1 square conditions (i.e., uncrowding occurs). As mentioned, uncrowding occurred for very few stimuli categories. In cases that didn't show uncrowding, the vernier representation deteriorated further when flankers were added (not shown). Note: the model outputs a cross-correlation quantifying how similar the model output is to the model output in the vernier alone condition (so the higher this cross-correlation, the better the performance). To make comparisons with human thresholds easier, we applied the same linking hypothesis as Hermens et al. (12): we fitted a psychometric function to link model outputs to behavioural results, as explained in the main text.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA6: Zhaoping's V1 recurrent model

Spatial extent	Mechanism	Organisation	Grouping component
Global	Pooling	Recurrent	No

This recurrent neural network model is described by Li Zhaoping (15). The network consists of a grid of neurons tuned to 12 orientations that are linked by lateral connections that follow a specific pattern (see Fig SA6a&b). The connectivity pattern allows the network to reproduce many experimental effects such as pop-out, figure-ground segmentation and border effects. It has also been shown to highlight certain parts of visual displays such as masked verniers (16), and we wondered if it could similarly produce uncrowding. We recoded the network from scratch following the detailed instructions and using the same parameters as in (15) and studied it as another recurrent model of early visual cortex. We ran all our stimuli and assessed performance by cross correlating each output with the output of the vernier without flankers. The magnitude of the cross-correlation is taken as a measure of vernier offset discrimination performance. The model produces crowding but not uncrowding consistently across the dataset (see Fig SA6c).

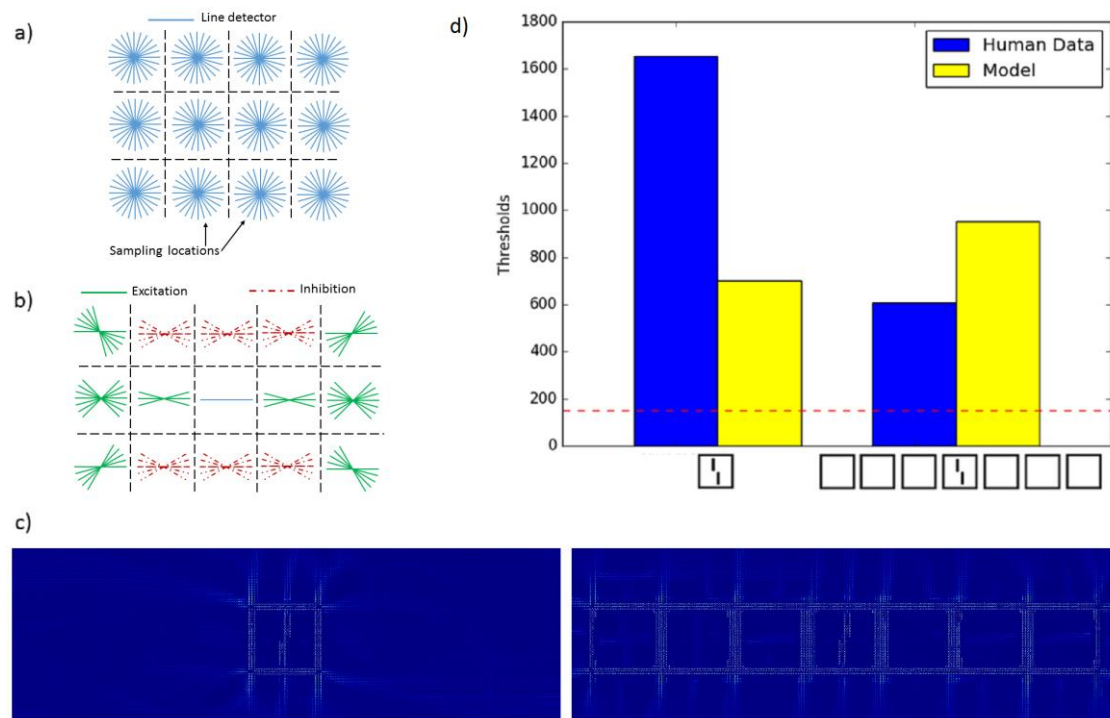


Fig SA6. V1 Segmentation model. **a.** The input is sampled at each grid position by neurons tuned to 12 orientations, mimicking V1 simple cells. **b.** The connectivity pattern between cells depends on their relative position and orientation as shown here. Solid lines indicate excitation and dashed lines indicate inhibition. As shown, each neuron excites aligned neurons and inhibits non-aligned neurons. Each neuron has the same connectivity pattern, suitably rotated and translated. **c.** Output images for the square category. Each small oriented bar shows the maximally active orientation at this grid position. **d.** Results for the squares category. The dashed red bar shows the vernier threshold, which is matched for humans and the model. As shown, uncrowding does not occur in the model, because performance is worse for the 7 squares than the 1 square stimulus. Note: the model outputs a cross-correlation quantifying how similar the model output is to the model output in the vernier alone condition (so the higher this cross-correlation, the better the performance). To make comparison with the human threshold easier, we applied the same procedure as we did for the epitomes, i.e., we applied the following monotonic transformation to the output: “threshold-like output” = $1/\text{“raw output”}$. Then we scaled the result to be in the same range as the human results. This monotonic re-scaling does not change the conclusions – the phenomenon of uncrowding cannot be altered.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA7: A variation of the Laminart model

Spatial extent	Mechanism	Organisation	Grouping component
Global	Pooling	Recurrent	Yes

The LAMINART model by Cao & Grossberg (17) is a neural network capable of computing illusory contours between collinear lines. Francis, Manassi & Herzog (18) augmented it with a segmentation process in which elements linked by illusory contours are grouped together by dedicated neural populations. This dedicated neural processing operates in the same way for all conditions and plays an important role in explaining many other visual phenomena (review: 44). This model process was intended as an implementation of a two-stage model of crowding, with a strong grouping process: stimuli are first segmented into different groups and, subsequently, elements within a group interfere. After dynamical processing, different groups are represented by distinct neural populations. Performance is determined by template matching. Importantly, crowding is low when the vernier is alone in its group (i.e., when the population representing the vernier does not represent other elements) and high otherwise.

The segmentation process is started by local selection signals and spreads along connected contours (Fig SA7). The location of each selection signal follows a Gaussian distribution centred on a given location, with a constant standard deviation. Uncrowding occurs when the selection signals hit a group of flankers without hitting the vernier, rescuing it from the deleterious effects of the flankers. In our simulations, each stimulus is run twenty times, each time drawing a new selection signal location. The final performance is averaged over these twenty trials. Crucially, segmentation becomes easier with more flankers, because a group of many flankers connected by illusory contours produces a larger region for selection (Fig SA7).

To account for the observers' proclivity to succeed in the vernier discrimination task, the central location of a selection signal is tuned to produce the least amount of crowding for any condition. This assumption follows the idea that an observer does the best job possible for any given situation. Although this added flexibility is not present in other models, it does not constitute an unfair advantage for the LAMINART. Indeed, it is not strictly necessary in order for the model to produce uncrowding. For example, if the segmentation signals' central location followed a uniform distribution over the whole stimulus, it would still hit a large group of flankers (without hitting the target) more easily than a small group of flankers. In summary, whenever the flankers form a wide group that can be easily segregated from the vernier, uncrowding should be produced. Hence, uncrowding is largely independent of the selection signals' distribution.

Many stimuli in the dataset had been simulated by the model in Francis et al. (18). Here, we improved the model by using more orientations and we ran the model on our full dataset, using the template matching measure (some stimuli could not be run for reasons detailed below). Overall, the LAMINART explains the data set well (Fig SA7). More precisely, the categories circles, Gestalts, lines, octagons, squares and hexagons are all well explained. Categories irreg1, irreg2 and stars cannot be explained, but they include bars of many different orientations, and the current LAMINART simulation is only capable of handling eight orientations. We did not run the stimuli in the patternStars and patternIrregular categories because they are too large to be processed in realistic time. In general, situations where the model fails tend to be those in which the model groups elements while the data suggests it should not, leading in some cases to no uncrowding, and in other cases to excessive uncrowding. One example is when flankers (e.g., squares and stars) group together when they should not. Another example is when flankers group with the target vernier (e.g., irreg1), suggesting the need to improve the grouping mechanism itself (Fig SA7). Across all stimuli and all models, the LAMINART is by far the most successful model in this comparative study because it can explain a wide range of uncrowding results, as well as capture classic crowding effects.

Back to Chapter 1: [\[Introduction\]](#) - [\[Methods\]](#) - [\[Results\]](#) - [\[Discussion\]](#)

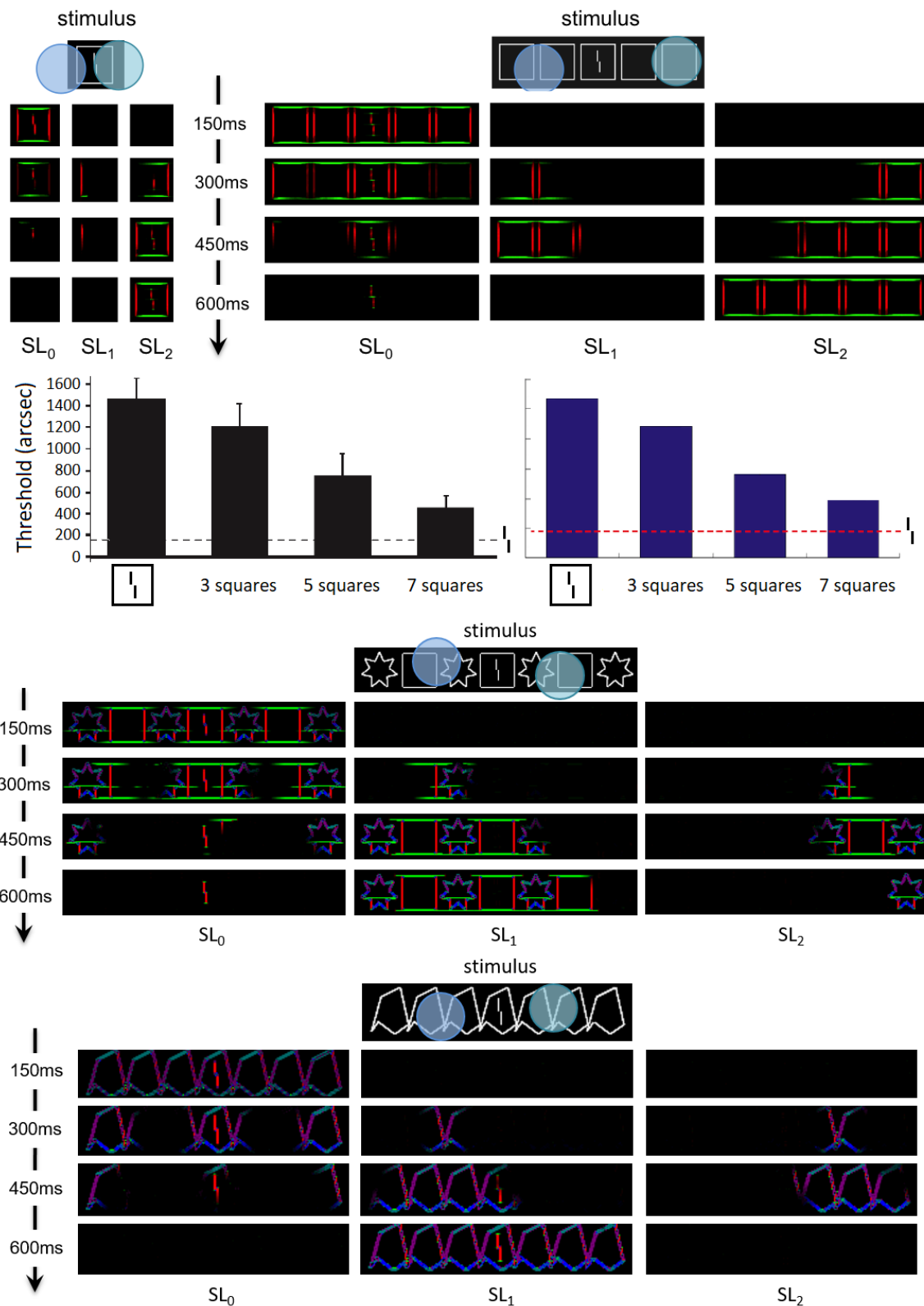


Fig SA7. The LAMINART variation. Top. Activity in the LAMINART model. Colors represent the most active orientation (red: vertical, green: horizontal). When a stimulus is presented, segmentation starts to propagate along connected (illusory or actual) contours from two locations marked by attentional selection signals. Visual elements linked together by illusory contours form a group. After dynamic, recurrent processing, the stimulus is

represented by three distinct neural populations, one for each group. Crowding is high if other elements are grouped in the same population as the vernier, and low if the vernier is alone. On the left, the flanker is hard to segment because of its proximity to the vernier. Across the trials, the selection signals often overlap with the whole stimulus, considered as a single group. Therefore, the flanker interferes with the vernier in most trials, and crowding is high. On the right, the flankers are linked by illusory contours and form a group that spans a large surface. In this case, segmentation signals can easily hit the flanker group successfully (without hitting the vernier). The vernier thus ends up alone in its group in most trials and crowding is low. **Top-center.** The left row shows human performance with the square flanker stimuli. The right row is the output of the LAMINART model. It fits the data very well. The same holds true for a majority of our stimuli. To compute the LAMINART's output values, we used the same linking hypothesis as in the original description of the model (18): template matching is used to decide if the target vernier offset is left or right, and this result is monotonically transformed into a threshold-like measure. **Bottom-center.** Sometimes flankers group together (illusory contours are formed) when they should not, and the model erroneously predict uncrowding for this condition. **Bottom.** Sometimes flankers group with the vernier when they should not. Here, weak illusory contours connect the central flanker and the vernier. No uncrowding can be produced for this condition because segmentation always spreads to the vernier, independently of the success of the selection signals.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA8: Alexnet (convolutional neural network)

Spatial extent	Mechanism	Organisation	Grouping component
Local	Pooling	Feedforward	No

Deep Convolutional Neural Networks (CNNs) are *local*, *feedforward*, *pooling* networks. Training involves using feedback signals to adjust weights between neurons in subsequent layers. Once the network has been trained, users typically fix the weights and use the network in a feedforward manner. Given enough time and training samples, CNNs can learn any function by learning adequate weights (20,21). CNNs fit very nicely in the standard view of vision research, in which basic features, such as edges, are combined in a hierarchical, feedforward manner to create higher-level representations of complex objects (Fig 2a in the main text of Chapter 1). We reasoned that crowding would occur in these networks for exactly the same reason as in classic local pooling models: the target and the flankers' representations at a given layer are pooled within the receptive fields of the subsequent layer, thus, leading to poorer performance. Although deep networks obviously compute groups such as objects or animals, these groups have no effect whatsoever on crowding of lower level features. Indeed, there are

no connections from higher to lower level layers. Thus, elements far away from the vernier cannot interact with nearby elements and lead to uncrowding. To test this hypothesis, we processed the square category through AlexNet (22), a deep net trained to classify natural images with high accuracy, using Tensorflow (23). In order to determine vernier offset discrimination in different layers, we trained classifiers to identify the vernier offset from the activations of different layers of Alexnet (Fig SA8a). The classifiers had a single hidden layer with 512 units, followed by a softmax layer with two outputs, corresponding to *left* and *right*. In the training phase, we ran verniers through the network, and trained classifiers to identify the offset orientation from the different layers' activations (which were normalized to zero mean and unit standard deviation). Each layer had its own classifier. We used all ReLU layers following the convolution layers and the last fully connected layer. A different classifier was trained for each of these layers. During the test phase, we used verniers alone, verniers flanked with a single square (crowded stimuli) and verniers with 7 squares flankers (uncrowded stimuli). Both training and testing stimuli had varying sizes, offsets and positions in the image. Fig SA8 shows average performance for each layer over 6 runs. For each run, we trained a new classifier on each layer, using 250000 verniers in the training set. In the testing phase, we ran 3000 verniers, 3000 crowded stimuli and 3000 uncrowded stimuli through Alexnet. Our classifiers identify vernier orientation from the layer activations for each of these inputs. Interestingly, our classifiers could well retrieve the test vernier orientations with 100% accuracy in all convolutional layers (layers 2, 3, 4 and 5). Adding square flankers deteriorated performance strongly. The single square (crowded) stimuli could be decoded only in the convolutional layers 2, 3 and 4, and in fully connected layer 7, but with much poorer accuracy than the vernier alone. Crucially, unlike in humans, the 7 squares (uncrowded) stimulus performance was always worse or equal to the performance on the single square (crowded) stimulus. Hence, the deep network produced crowding, but not uncrowding. We suggest that the mechanism leading to these results is similar to the classic local pooling account of crowding.

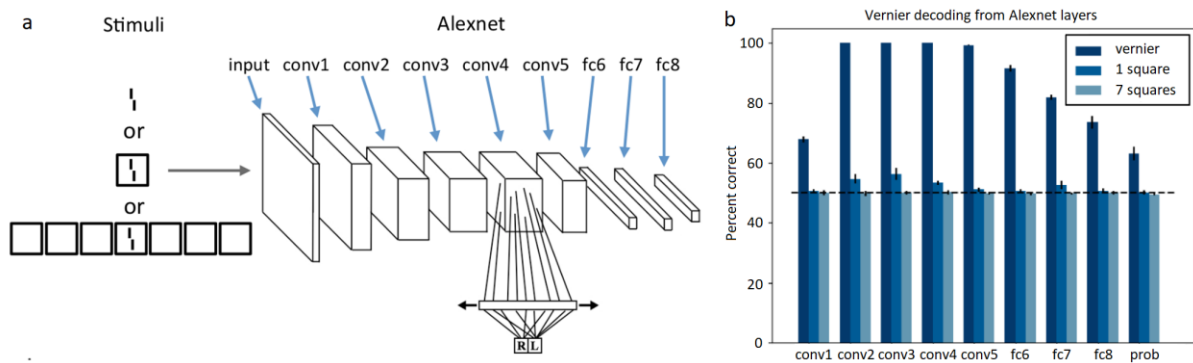


Fig SA8. Alexnet. **a.** Stimuli consisted of either verniers, verniers surrounded by a single square or verniers with seven squares. The stimuli had varying sizes, vernier offsets and positions. Alexnet's architecture and a classifier are shown on the right (there was a classifier at each layer). The boxes correspond to the input (leftmost box) and activated neuron layers (see 47 for the detailed architecture of Alexnet). We trained softmax classifiers on all ReLU layers following the convolution layers and the last fully connected layer to detect vernier orientation from the layer's activity. **b.** Accuracy of softmax classifiers trained to detect vernier orientation from different layers in the deep neural network Alexnet. Across all layers, the offsets in crowded stimuli (1 square flanker) are always better detected than offsets in uncrowded stimuli (7 square flankers). This runs contrary to human performance. NB. This model only produces percent correct, there is no output image.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA9: Hierarchical sparse selection (HSS)

Spatial extent	Mechanism	Organisation	Grouping component
Local	Sparse readout	Feedforward/recurrent	No

This model was described by Chaney, Fischer & Whitney (24). In a series of experiments, it was shown that in spite of difficulty identifying a crowded target, crowding does preserve some information about the target, i.e., information is rendered inaccessible but not destroyed (see 7,8 for reviews). For example, a face surrounded by other faces cannot be explicitly identified, but information about its features can nevertheless survive crowding and contribute its holistic attributes to the perceived average of a set of faces (27). To accommodate these results, Chaney et al. (2014) proposed that information is not lost along the visual processing hierarchy. Instead, crowding occurs because readout is sparse. Specifically, given a feature map representing a stimulus, only a subset of the neurons from this map can be used to decode the target, which leads to crowding's deleterious effects (Fig SA9a).

Using the author's code, we tested all our stimuli and found that crowding could be explained, but uncrowding did not occur in the model (Fig SA9b). Originally, the model was used to detect crosses, triangles and circles. We modified the model's readout layer to classify vernier orientation, which was achieved with 99.13% accuracy (the rest of the model does not need any change to accommodate new stimuli). Then, we dropped 75% of the neurons for the imperfect readout, which led to a vernier classification accuracy of 81.48%. We tested all our stimuli by asking the model to classify the vernier orientation, first without dropping any neurons, then with 75% of the neurons dropped for the sparse readout, as we did for the verniers. For all stimuli, performance dropped with the sparse readout. For example, the 1 square condition was classified with 93.35% accuracy when all neurons were used, and this dropped to 75.55% with sparse readout. The 7 squares condition had a similar profile, but classification accuracy was worse than for the 1 square condition (71.73% with all neurons and 59.23% with sparse readout). This pattern of results was found in all stimulus categories: sparse readout impaired performance and adding more flankers impaired performance too. Thus, there was crowding but no uncrowding. We would like to mention that Chaney et al. argue that uncrowding can in fact be explained, if the target and flanker are represented in different feature maps, which are however not implemented at the moment. In essence, visual stimuli are segmented into different feature maps (this must happen early in the visual pathway to explain the low-level vernier results), and subsequently the HSS model applies within feature maps, on this pre-segmented input.

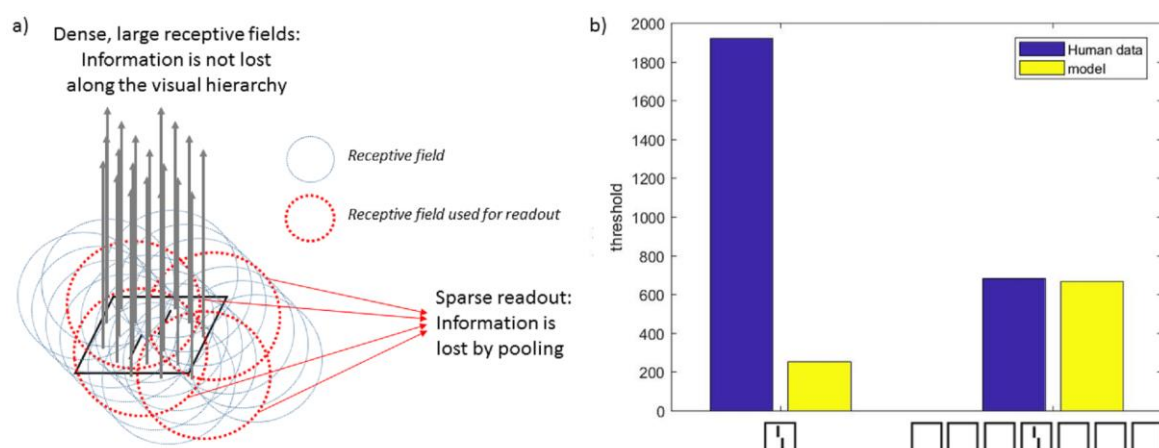


Fig SA9. Hierarchical Sparse Selection model. **a.** The model posits that receptive fields along the visual hierarchy are large and dense. This allows for “lossless” transmission of information through the visual system. For instance, the offset of the vernier in this illustration is not corrupted by pooling thanks to the density of the receptive fields

(blue and red circles). Crowding occurs because, when we try to access information, only a few sparse receptive fields are used for readout (red circles). Hence, crowding occurs at readout because of sparse sampling of receptive fields. This sparse readout can occur at any stage of visual processing, for example from low-level features (shown here) to faces. **b.** Uncrowding does not occur in the Hierarchical Sparse Selection model because performance is worse for the model on the 7 squares than the 1 square condition, contrary to human performance. NB. This model only produces a scalar output, there is no output image.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA10: Saccade-confounded summary statistics

Spatial extent	Mechanism	Organisation	Grouping component
Local	Pooling	Feedforward	No

Nandy & Tjan (28) proposed a model linking summary statistics to saccadic eye movements: crowding is proposed to occur because the acquisition of summary statistics in the periphery is confounded by eye-movement artifacts. This leads to inappropriate contextual interactions in the periphery and in this way produces crowding. For the present purposes this is not directly relevant, because foveal and peripheral uncrowding results are qualitatively identical (29), which the saccade-confounded summary statistics model cannot explain since it suggests that crowding can only occur in peripheral regions. Moreover, it is not clear how uncrowding can occur in this model.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA11: Population coding

Spatial extent	Mechanism	Organisation	Grouping component
Local	Overlapping population codes	Feedforward	No

This kind of model was first described by Van den Berg, Roerdink, & Cornelissen (30). A similar model was proposed by Harrison & Bex (31). Both models elegantly produce both pooling and substitution behaviour by assuming that an element's orientation is represented by a population code: a probability distribution of its orientation. When many elements are present, the population codes interfere and disturb the target element's representation, which leads to

crowding. This interference depends on distance and is usually modeled as a 2D Gaussian. Dayan & Solomon (32) proposed a model in which elements are represented as probability distributions. They added a Bayesian process to account for the accumulation of evidence over time. Their model captures local crowding effects similarly to Van den Berg et al. and Harrison & Bex's models: the interference comes from the representations of neighbouring elements deleteriously affecting each other. This model and the one by Van den Berg and colleagues cannot handle images as input and thus could not be tested with our stimuli.

We have shown elsewhere that the Harrison & Bex (31) implementation cannot explain uncrowding (33). Agaoglu & Chung (34) showed that the interaction between elements depends on which of them is considered as the target for report. Hence, the crowding interference between elements in the display depends on the task, which is not easily incorporated in the models without a dedicated process. Van den Berg et al. (30) suggested that elements do not interfere when they are represented in different perceptual groups, similar to the LAMINART model. Similarly, Harisson & Bex (31) have suggested that a preprocessing stage determining which elements interfere is needed.

Back to Chapter 1: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SA12: Fourier model

Spatial extent	Mechanism	Organisation	Grouping component
Global	Fourier	Feedforward	No

The Fourier transform is sensitive to global aspects of spatial configurations because it is based on periodic features. Even if it was never explicitly proposed to explain crowding, it may capture some effects of uncrowding that have to do with regularities in the stimulus. Previously (14,35), we used a Fourier-based model and tested it on the entire dataset. Essentially, this is a texture-like model, assuming that the brain Fourier transforms the visual input. Repetitive structures, such as arrays of squares are more compactly coded in the Fourier space than the 2D space. We restate the results here for comparison with the other models. The model first bandpass filters the stimuli (passing a small range of frequencies at all orientations), then computes the Fourier transforms of the filtered left- and right-offset cases for each stimulus. Similarly to what was done to measure performance of Zhaoping's recurrent V1 model, these

are cross-correlated with the filtered versions of the verniers without any flankers and the magnitude of the cross-correlation is taken as a measure of vernier offset discrimination performance. This process is repeated over all possible passbands (which is finite given a fixed image size) until the pass-band yielding performance most similar to humans is found. Across the dataset, this approach failed to reproduce the data (see Fig SA12), suggesting that such a simple use of global regularities in the display is insufficient to explain crowding. Depending on the set of Gabor filters, uncrowding occurred for certain stimuli, but this was never consistent over several stimulus types, which is suggestive of overfitting. With one set of filters the lines category could be explained, with another the Gestalts category could be explained.

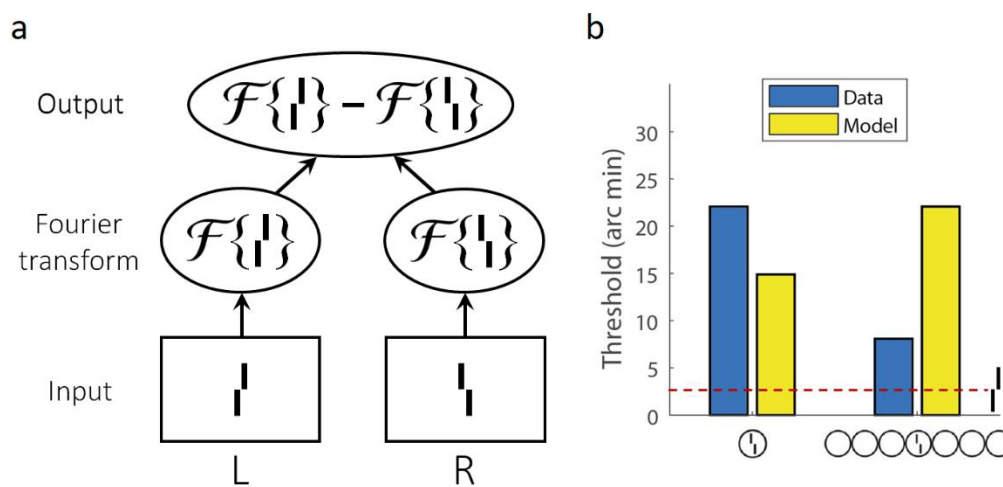


Fig SA12. Fourier model. **a.** The Fourier model computes Fourier transforms for the left- and right-offset versions of each stimulus. If these transforms are very different, crowding is low because the offset direction is easy to decode in Fourier space (35). **b.** Output of the Fourier model. The black bars represent human data; the white bars represent the model output. The model failed on most stimuli (35). Note: this model only produces a scalar output, there is no output image.

Back to Chapter 1: [\[Introduction\]](#) - [\[Methods\]](#) - [\[Results\]](#) - [\[Discussion\]](#)

B: Supplementary information for Chapter 2

Back to Chapter 2: [[Introduction](#)] - [[General Materials and Methods](#)] - [[TTM & Grouping Effects](#)] - [[TTM & Face Crowding](#)] - [[Discussion](#)]

SB1: Comparison between Lines and Completion experiments

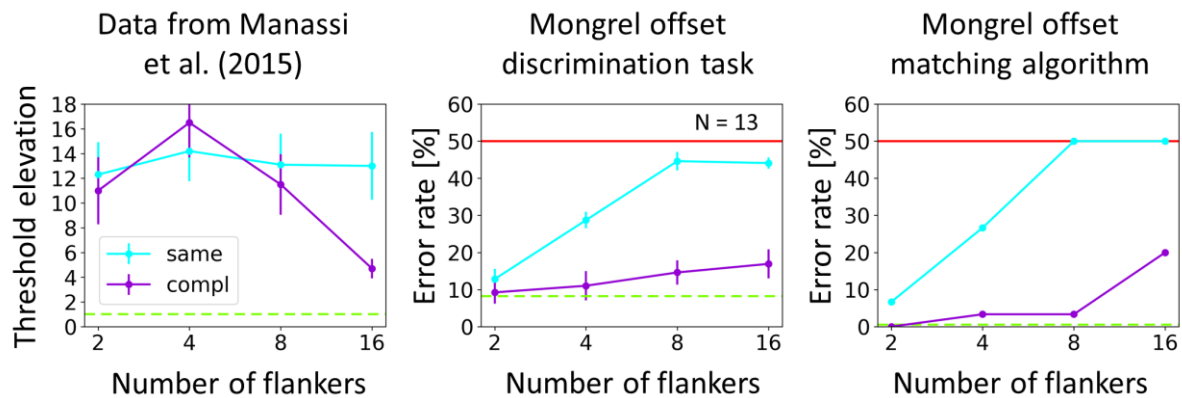


Figure A. Comparison between Lines and Completion experiments. **Left.** Offset discrimination thresholds were determined for vernier targets presented in peripheral vision. Bars indicate flanker configurations threshold elevation compared to the vernier alone (green dashed line). Data are taken from Manassi et al. (2015). **Center.** As a validation of the TTM, we asked observers to discriminate between left and right offset vernier in mongrel images. Green dashed line indicates vernier alone performance. Red line indicates chance level (50% accuracy). **Right.** As a further model validation, we measured the performance of our template matching algorithm, using the same mongrels as in the human experiment. We compared the crowding induced by different number of same length flankers, with (same, blue) and without (compl, purple) the mask. In both our validation tasks, crowding was always weaker with than without the mask, contrary to the human data, in which this effect appears only for 16 flankers. Moreover, with or without adding the mask, crowding always increased with more flankers.

Back to Chapter 2: [[Introduction](#)] - [[General Materials and Methods](#)] - [[TTM & Grouping Effects](#)] - [[TTM & Face Crowding](#)] - [[Discussion](#)]

SB2: Shapes experiment with diamonds

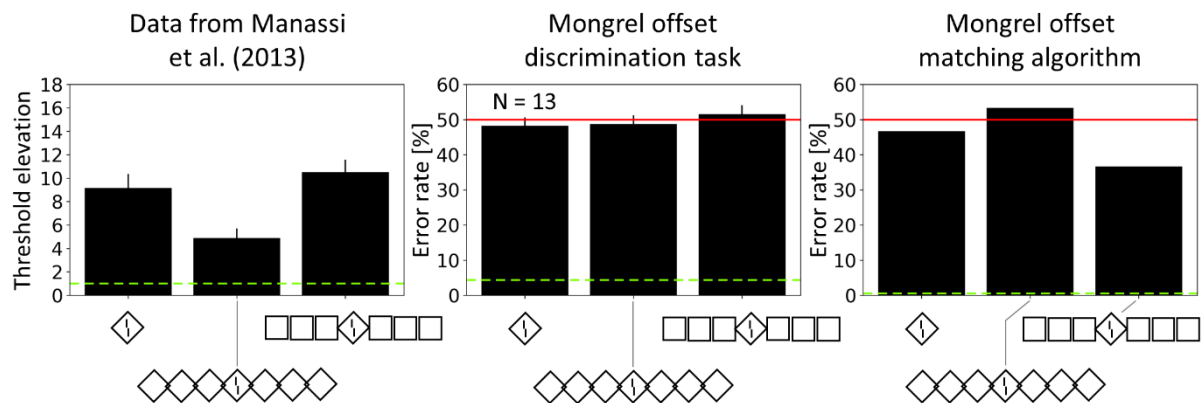


Figure B. Shapes experiment with diamonds. **Left.** Offset discrimination thresholds were determined for vernier targets presented in peripheral vision. Bars indicate flanker configurations threshold elevation compared to the vernier alone (green dashed line). Data are taken from Manassi et al. (2013). **Center.** As a validation of the TTM, we asked observers to discriminate between left and right offset vernier in mongrel images. Green dashed line indicates vernier alone performance. Red line indicates chance level (50% accuracy). **Right.** As a further model validation, we measured the performance of our template matching algorithm, using the same mongrels as in the human experiment. In the original experiment, crowding was strong when the vernier target was flanked by a single diamond and decreased when three additional diamonds were added on each side (1st column, 1D vs 7D). When the flanking diamonds were rotated by 45°, crowding was strong again (1st column, 6S1D). The TTM did not reproduce this set of results: for both our model validation tasks (2nd and 3rd columns) crowding was strong for all tested conditions, independently of the flanker configuration. The same validation was performed with a different fovea radius parameter in the TTM, yielding similar results ([Suppl. Inf. B, SB3](#)).

Back to Chapter 2: [[Introduction](#)] - [[General Materials and Methods](#)] - [[TTM & Grouping Effects](#)] - [[TTM & Face Crowding](#)] - [[Discussion](#)]

SB3: Shapes and Patterns experiments with larger fovea parameter

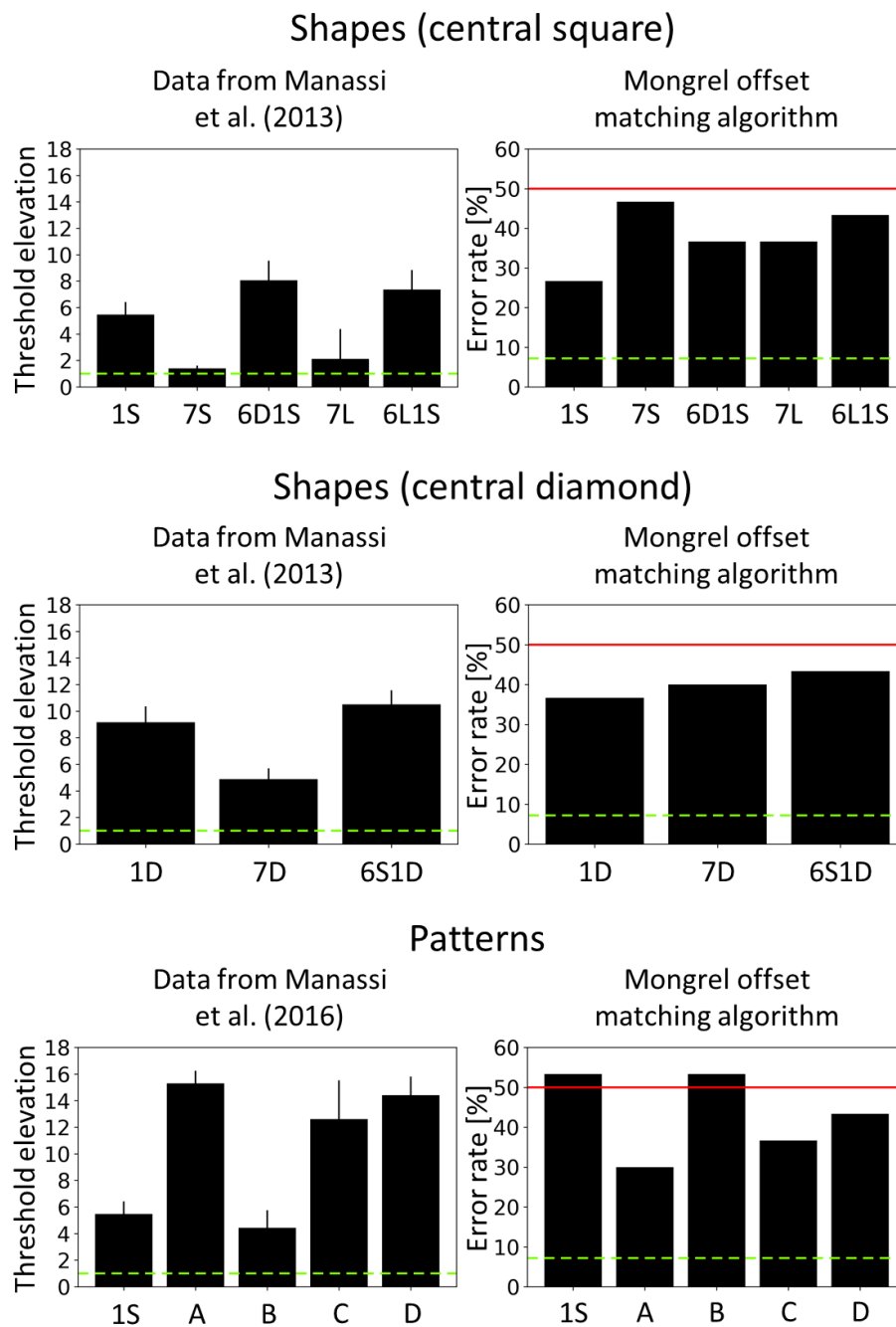


Figure C. Shapes and Patterns experiments with a larger fovea radius parameter in the TTM. **Left.** Offset discrimination thresholds were determined for vernier targets presented in peripheral vision. Bars indicate flanker configurations threshold elevation compared to the vernier alone (green dashed line). Data are taken from Manassi et al. (2013). **Right.** As a validation of the TTM, we measured the performance of our template matching algorithm. Results are qualitatively similar to the ones depicted in Figure 3 and in [Suppl. Inf. B, SB2](#).

Back to Chapter 2: [\[Introduction\]](#) - [\[General Materials and Methods\]](#) - [\[TTM & Grouping Effects\]](#) - [\[TTM & Face Crowding\]](#) - [\[Discussion\]](#)

SB4: Butterflies experiment

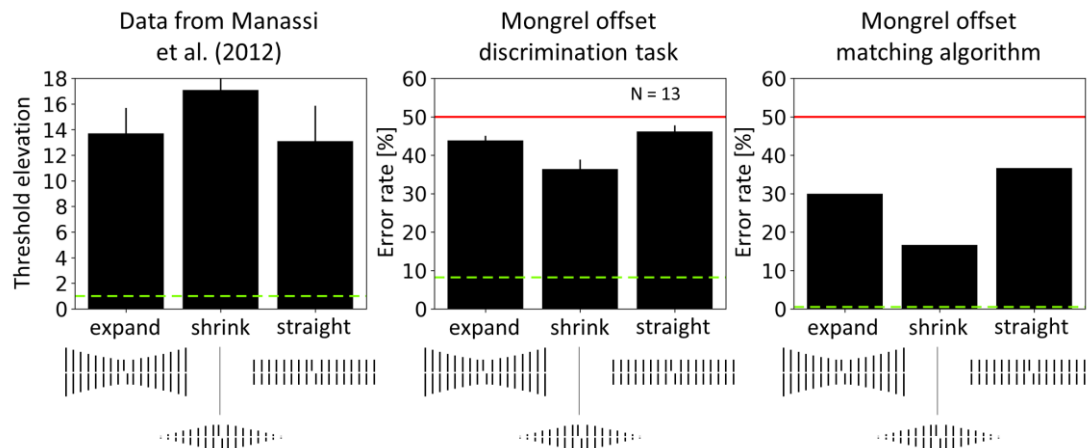


Figure D. Butterflies experiment. **Left.** Data from (Manassi et al., 2015). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. **Center.** TTM validation in which observers discriminate between left and right offset verniers in mongrel images. **Right.** TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy).

Back to Chapter 2: [\[Introduction\]](#) - [\[General Materials and Methods\]](#) - [\[TTM & Grouping Effects\]](#) - [\[TTM & Face Crowding\]](#) - [\[Discussion\]](#)

SB5: TTM and prediction power - Template match algorithm performance

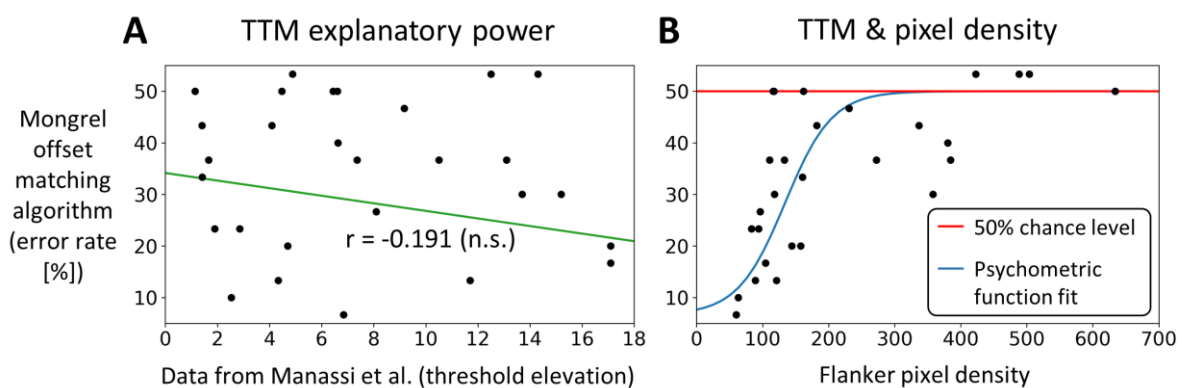


Figure E. A. When plotting error rates in the mongrel offset matching algorithm as a function of psychophysical thresholds data from Manassi et al. (2012, 2013, 2015, 2016), no correlation was found ($r(36)=-0.191$, $p=-0.264$, $BF_{01}=2.647$). **B.** We plotted the error rates measured in the mongrel offset matching algorithm with all tested flanking conditions as a function of the sum of the flanker pixel density (see Methods for details). Each dot indicates a flanking condition in Figure 1. The red line indicates chance level performance. The data are well fitted by a psychometric function (blue line, see Method for details). The correlation between the measured error rates and the error rates predicted by the fitted function is strong ($r(36)=0.739$, $p<0.001$, $BF_{10}>10^4$).

Back to Chapter 2: [\[Introduction\]](#) - [\[General Materials and Methods\]](#) - [\[TTM & Grouping Effects\]](#) - [\[TTM & Face Crowding\]](#) - [\[Discussion\]](#)

SB6: TTM and prediction power - Separate experiments

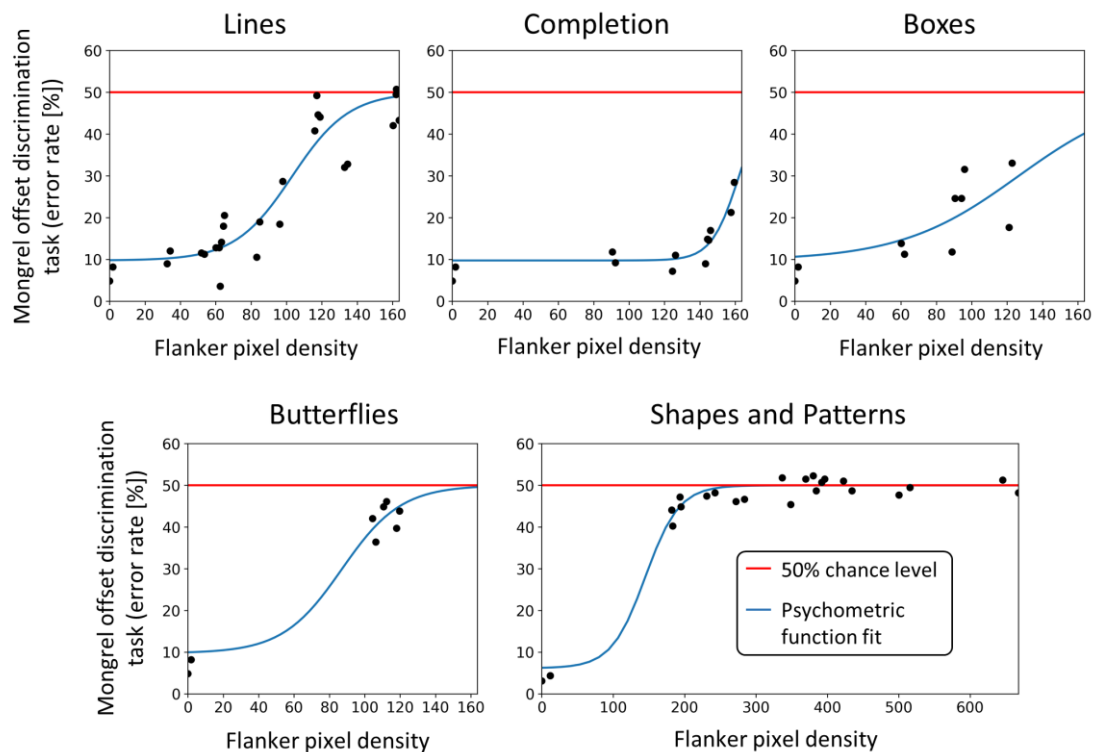


Figure F. We plotted the error rates measured in the mongrel offset discrimination task with all tested flanking conditions as a function of the sum of the flanker pixel density (see Methods for details). Each dot indicates a flanking condition in Figure 1. The red line indicates chance level performance. We fitted the datapoints for all experiments separately, using a psychometric function (blue lines, see Method for details). To have more datapoints in the fits, we also used the conditions in which we removed the pointers. The correlations are all significant. “Lines”: $r(24)=0.929$, $p<0.001$; “Completion”: $r(10)=0.923$, $p<0.001$; “Butterflies”: $r(6)=0.981$, $p<0.001$; “Boxes”: $r(8)=0.759$, $p=0.011$; “Patterns” and “Squares”: $r(22)=0.992$, $p<0.01$.

Back to Chapter 2: [\[Introduction\]](#) - [\[General Materials and Methods\]](#) - [\[TTM & Grouping Effects\]](#) - [\[TTM & Face Crowding\]](#) - [\[Discussion\]](#)

SB7: Pointers location in Manassi et. al (2012)

Stimuli as depicted in Rosenholtz et al. (2019)



Stimuli used in Manassi et al. (2012)

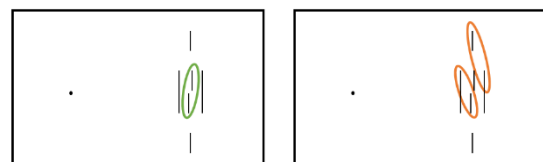


Figure G. In Manassi et al. (2012, 2013, 2015), pointers were added above and below the target to reduce its location uncertainty. It was argued that these pointers may increase crowding by creating multiple offsets among vernier, flankers and pointers lines (Rosenholtz et al., 2019) (left). However, the pointers used in the actual experiment were further from the vernier than reported by the authors (right), reducing the likelihood that pointers create more crowding.

Back to Chapter 2: [[Introduction](#)] - [[General Materials and Methods](#)] - [[TTM & Grouping Effects](#)] - [[TTM & Face Crowding](#)] - [[Discussion](#)]

SB8: Effect of pointers in the TTM

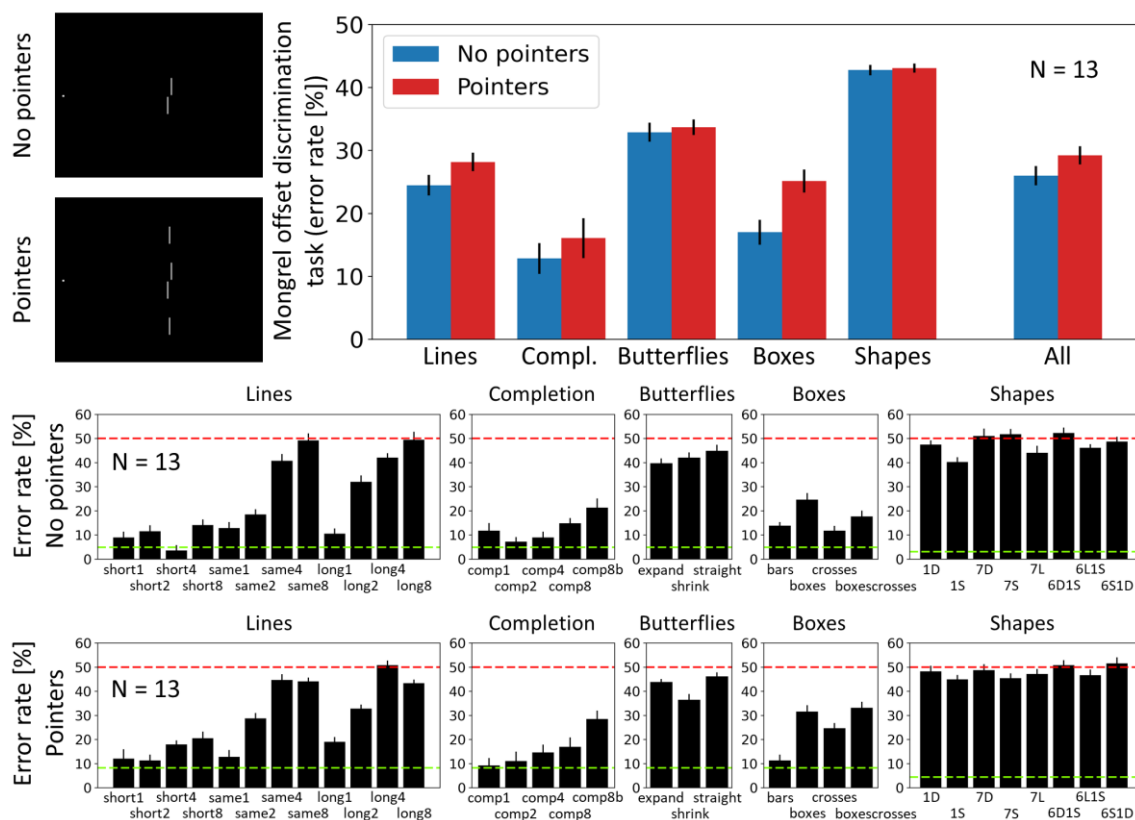


Figure H. We measured human performance in the mongrel offset discrimination task for all conditions in Manassi et al. (2012, 2013, 2015), with or without pointers (bottom). The actual layout of the different conditions is shown in Figure 1. The TTM did not show any significant increase in crowding strength (top panel, “All”, $t(12)=1.485$, $p=0.151$). Analyzing the conditions separately, not correcting for multiple comparisons to maximize evidence for an effect of pointers, only the “Boxes” experiment exhibited a significant difference ($t(12)=2.905$, $p\text{-value}=0.008$). All the other conditions did not (“Lines”: $t(12)=1.162$, $p=0.119$; “Completion”: $t(12)=0.776$, $p=0.445$; “Butterflies”: $t(12)=0.382$, $p=0.706$, “Shapes”: $t(12)=0.273$, $p=0.787$).

Back to Chapter 2: [[Introduction](#)] - [[General Materials and Methods](#)] - [[TTM & Grouping Effects](#)] - [[TTM & Face Crowding](#)] - [[Discussion](#)]

SB9: Single face discrimination task - reverted back

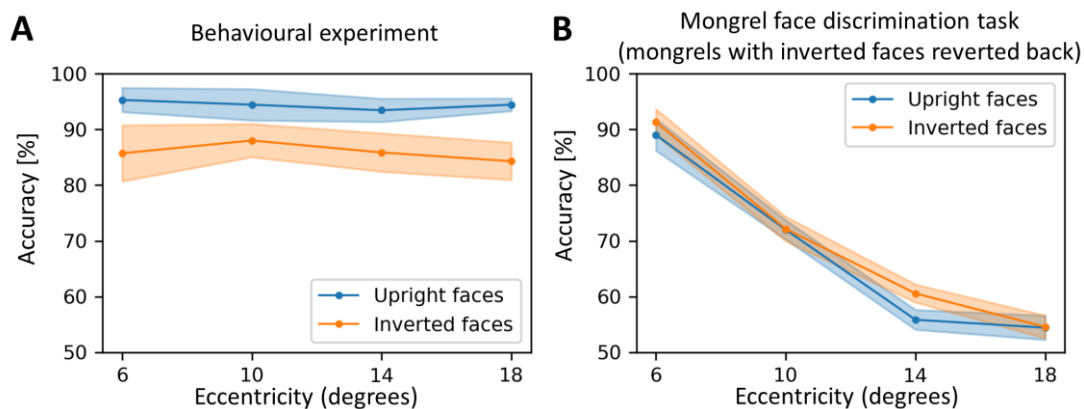


Figure I. TTM & single Mooney faces, reverted-back version. **A.** Single face discrimination task. Observers were able to discriminate an upright or an inverted face from a scrambled face at all tested eccentricities. Moreover, performance was higher for upright than for inverted faces. **B.** Mongrel single face discrimination task. In this task, the mongrels that came from original stimuli in which the face was inverted were reverted back, so that they appeared upright to the observers. This was done in order to isolate inversion effects in the TTM from inversion effects in humans as much as possible. As in Figure 7B, performance decreased when the eccentricity was increased, contrary to the behavioral results. Moreover, no significant difference between the upright and inverted face conditions was observed. The data were analyzed using a linear mixed effect model, with eccentricity and face orientation as the two fixed effects and individual subjects as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1)=0.015$, $p=0.902$). The main effect of eccentricity was significant ($\chi^2(1)=94.862$, $p<0.001$), but the effect of face orientation was not ($\chi^2(1)=1.158$, $p=0.282$). The difference in effect size between the full model, including both effects and the reduced model excluding the effect of face orientation, was only 0.4% (full model: $r_m^2=0.819$, $r_c^2=0.826$, reduced model: $r_m^2=0.815$, $r_c^2=0.822$).

Back to Chapter 2: [[Introduction](#)] - [[General Materials and Methods](#)] - [[TTM & Grouping Effects](#)] - [[TTM & Face Crowding](#)] - [[Discussion](#)]

SB10: Mongrel gender matching algorithm

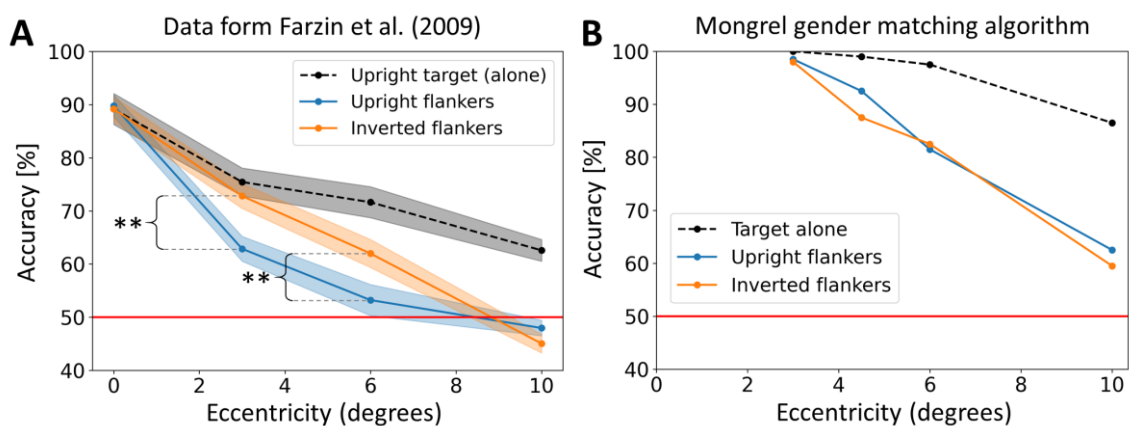


Figure J. TTM & crowding in Mooney faces, algorithm results. **A.** Face crowding task, data from Farzin et al. (2009). Target discrimination performance decreased when eccentricity increased. When the target face was flanked by inverted faces, crowding increased with increasing eccentricity (orange). When the target was flanked by upright faces, crowding increased even more with eccentricity (blue). **B.** Mongrel gender matching algorithm results. As with the gender crowding discrimination task, accuracy decreased with eccentricity but did not differ between the upright and inverted flanker conditions, contrary to the behavioural data.

Back to Chapter 2: [\[Introduction\]](#) - [\[General materials and methods\]](#) - [\[TTM & grouping effects\]](#) - [\[TTM & face crowding\]](#) - [\[Discussion\]](#)

SB11: TTM & Pixel density - Detailed methods

To assess the behaviour of the TTM, we plotted human performance in the mongrel vernier offset discrimination task (error rate [%]) against the flanker pixel density in the original stimuli. For each stimulus image, the flanker pixel density was computed as a weighted sum of the pixels that belong to the flanking pattern. Each pixel contribution was weighted by a function that decreased exponentially with the distance to the target (Eq. 2), mimicking Bouma's law (Bouma, 1970).

$$S = \sum_{i,j} e^{-D(i,j)^2/\sigma^2} \quad (2)$$

S was the sum of all pixel contributions, $D(i, j)$ the distance from pixel (i, j) to the target and σ the width of the weighting function. σ was set to the target eccentricity divided by 4 so that weights vanished for distances bigger than Bouma's law radius. To evaluate how close the TTM was to a simple pooling model, we fitted a psychometric function to the TTM performance (Eq. 3).

$$P(S \mid a, b, c) = 100 \cdot [\tanh(a \cdot S - b) \cdot (0.5 - c) + c] \quad (3)$$

P was the output performance (error rate [%]) computed by the fitted psychometric function, a , b and c were the fitted parameters. P was bounded by a basic error rate (c) and chance level (50%).

Back to Chapter 2: [\[Introduction\]](#) - [\[General materials and methods\]](#) - [\[TTM & grouping effects\]](#) - [\[TTM & face crowding\]](#) - [\[Discussion\]](#)

C: Supplementary information for Chapter 4

Back to Chapter 4: [\[Introduction\]](#) - [\[Methods\]](#) - [\[Results\]](#) - [\[Discussion\]](#)

SC1: Bouma's law model

To set a basis for our analysis, we used a model that assumes Bouma's law (42) holds true in dense displays. In this model (Fig SC1, 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 the main text of Chapter 4, 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.

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

$P_{unflanked}$ comes from the sparse display experiment in Van der Burg et al. (43) 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. (43), 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).

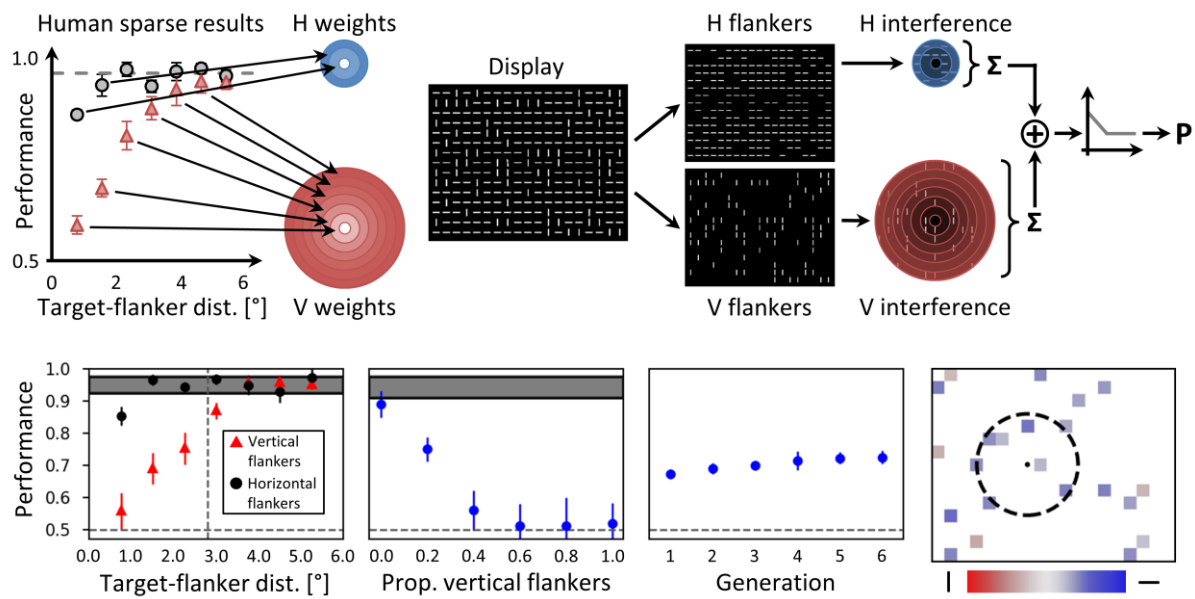


Fig SC1. Top. Bouma model. Flanker-target distance-dependent weights were defined as how much performance dropped from the unflanked level in the sparse display experiment of Van der Burg et al. (43). For each display, the probability of correct response is a decreasing function of the sum of its flankers' 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 of Chapter 4).

The results for both ways of selecting the best displays between generations are shown in Fig 3 in the main text of Chapter 4 (2nd row) for the first version and in Fig SC1 (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 preference 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.

Back to Chapter 4: [\[Introduction\]](#) - [\[Methods\]](#) - [\[Results\]](#) - [\[Discussion\]](#)

SC2: Population coding model

The population coding model (44) 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 SC2). 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 (44), 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 $[-45^\circ, 45^\circ]$, 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 $\pm 10^\circ$ (instead of $\pm 5^\circ$), for dense displays only.

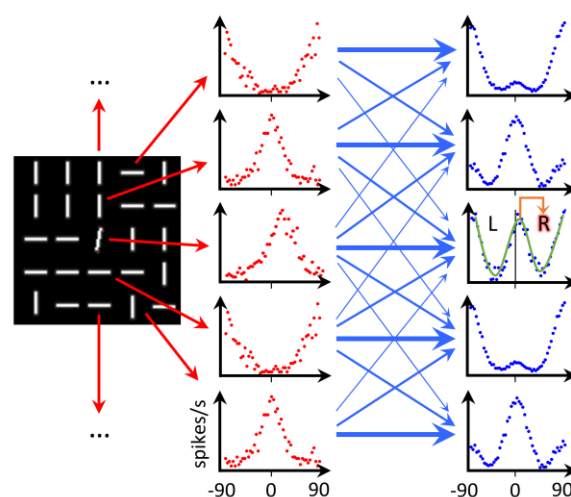


Fig SC2. Population coding model. Populations of orientation-selective neurons encode the content of every element in the display array (red arrows). Then, the activity responsible for each location of the array is pooled to a second layer of neuron populations (blue arrows). Pooling weights (represented here by the thickness of the

arrows) depend on the cortical distances between the populations. 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 of Chapter 4 (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 preference 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 (45) 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 preference measure.

Back to Chapter 4: [\[Introduction\]](#) - [\[Methods\]](#) - [\[Results\]](#) - [\[Discussion\]](#)

SC3: Texture model

Texture models (46) 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. (47) proposed that crowding is the result of such statistics being computed over pooling regions. They proposed to use the model of Portilla et Simoncelli (46) 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. (40) 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. (43). 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 SC3a). 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 SC3b).

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 $\pm 15^\circ$ (instead of $\pm 5^\circ$), 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 SC3b.

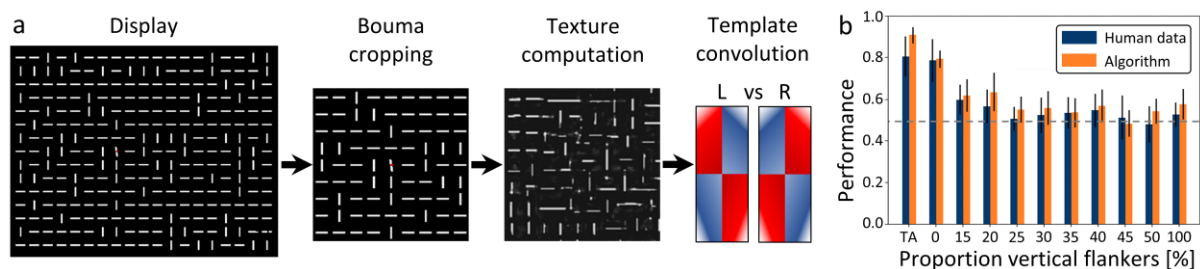


Fig SC3. a. Texture model. First, the stimulus display is cropped, so that only a Bouma-sized patch around the target is sent to the texture model. Then, the model iteratively matches a set of statistics between the input patch

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 of Chapter 4 (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 preference measure did not highlight any location, exactly as with the second version of the Bouma model (Fig SC1, bottom). In summary, the texture model only reproduced human results for the proportion measure.

Back to Chapter 4: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SC4: CNN classifier

Deep feedforward convolutional neural networks (CNNs) share many similarities with humans in their architecture, in their activity patterns (48,49), as well as in the performance they reach in a large number of visual tasks (50,51). Here, we used the same method as in (52), testing AlexNet (53) as a representative of CNNs, because it is often used as a model of the human visual system (54–57). The weights of AlexNet were already trained on ImageNet (58). 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 SC4). 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.

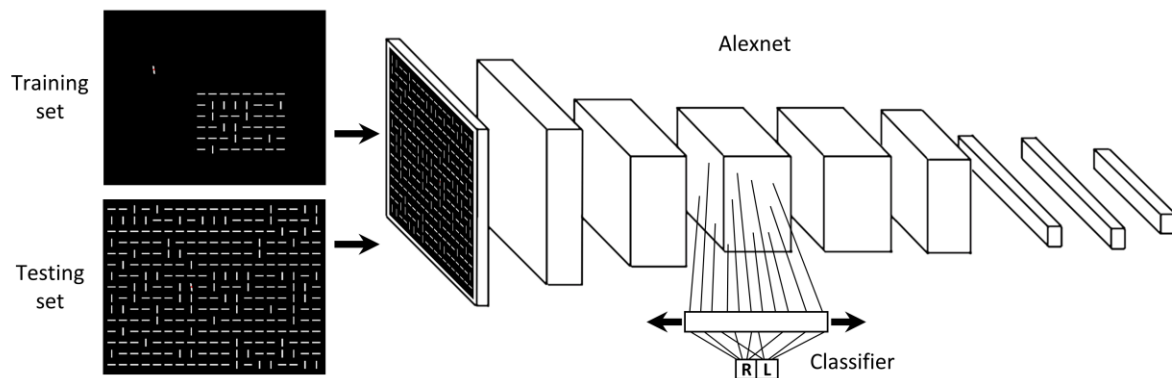


Fig SC4. CNN classifier. The stimulus display is processed by the architecture of Alexnet. On top of each layer, a decoder was trained to discriminate between left or right targets from the layer activity. The weights of Alexnet (which have been previously trained on ImageNet) did not change during the 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 (52).

The results obtained with the model are shown in Fig 3 in the main text of Chapter 4 (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. 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 preference measure did not highlight any specific location that was crucial for this improvement. In summary, the CNN classifier replicated none of the human results.

Back to Chapter 4: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SC5: Contour segmentation model (“Laminart”)

The Laminart model (59) 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 SC5). 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. (43). For each display, the algorithm links neighbouring bars whenever they or their tips are aligned (Fig SC5, 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. (59), 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 Suppl. Inf.. 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.

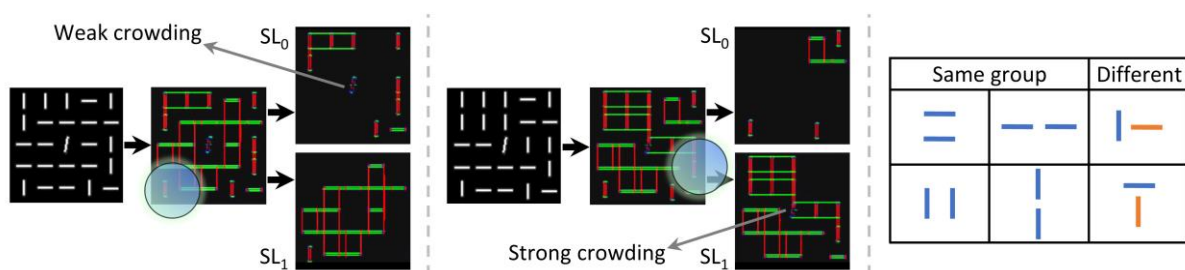


Fig SC5. Laminart model. In the original model (59), the stimulus is processed by an array of orientation-selective feature detectors. Coloured pixels that are depicted in the images correspond to the most active oriented cell at that location (red, green, blue, purple and turquoise for vertical, horizontal, 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. (43). 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.

Results obtained with the Laminart model are shown in Fig 3 in the main text of Chapter 4 (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 preference 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 SC5a, 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.

Back to Chapter 4: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SC6: Capsule network

Capsule networks are deep neural networks in which layers of neurons communicate through a recurrent process that implements grouping (Fig SC6). 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" (60), 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 SC6, right). The entire network is trained end to end through backpropagation. Doerig et al. (61) 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. (61). 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 SC6, 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.

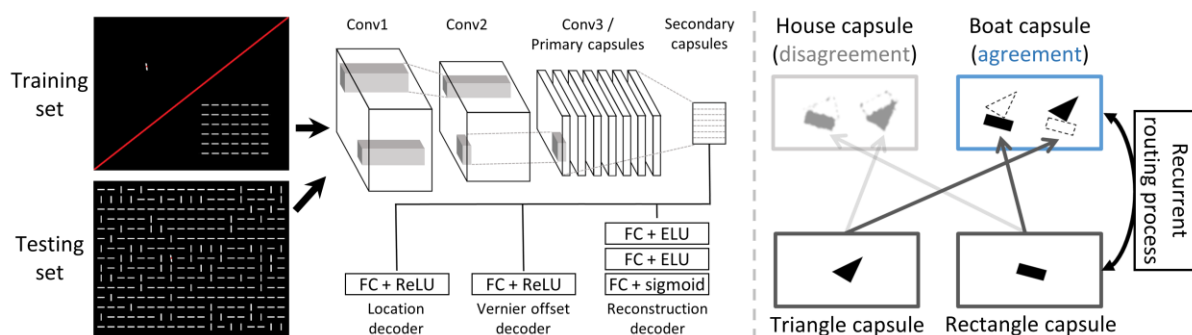


Fig SC6. Left. Capsule network. The stimulus display is processed by a set of convolutional layers, conveying information to the primary capsules, which then projects to the secondary capsules. Routing by agreement happens between the primary and secondary capsules. In this case primary capsules encode 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 (61).

Results obtained with the Capsule network are shown in Fig 3 in the main text of Chapter 4 (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 preference 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 preference measure. One caveat is that only one out of the 10 models we trained reached good target discrimination in dense displays.

Back to Chapter 4: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SC7: 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. (43) for sparse displays. On the other hand, the population coding model (44) is the best at explaining sparse display results but does not replicate the preference measure. For these reasons, we combined both models into a two-stage model (Fig SC7).

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.

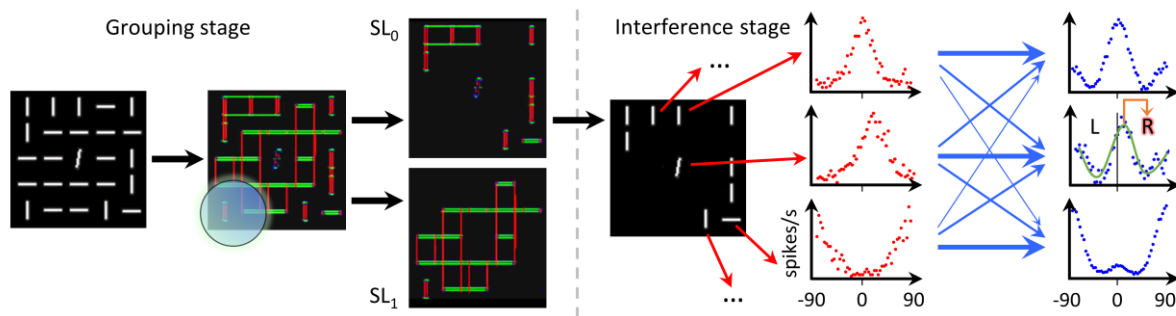


Fig SC7. Popart model. The model is composed of two stages. **Left.** Grouping stage. The Laminart model algorithm is used to parse the stimulus in different segmentation layers. **Right.** Interference stage. From the output of the segmentation algorithm, a new stimulus is built. Only the elements present 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 of Chapter 4 (last row). Thanks to the combination of both segmentation and population coding models, the Popart model qualitatively reproduces human results for all measures.

Back to Chapter 4: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

SC8: 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 (62). 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. (43), 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 of Chapter 4). The target was always displayed at 6° of eccentricity (8^{th} row, 8^{th} 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 of Chapter 4 (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).

Back to Chapter 4: [[Introduction](#)] - [[Methods](#)] - [[Results](#)] - [[Discussion](#)]

References

1. Jojic N, Frey BJ, Kannan A. Epitomic analysis of appearance and shape. In: Proceedings Ninth IEEE International Conference on Computer Vision. 2003. p. 34–41 vol.1.
2. Portilla J, Simoncelli EP. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *Int J Comput Vis.* 2000 Oct 1;40(1):49–70.
3. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis.* 2009 Nov 1;9(12):13–13.
4. Zhang X, Huang J, Yigit-Elliott S, Rosenholtz R. Cube search, revisited. *J Vis.* 2015;15(3):9–9.
5. Freeman J, Simoncelli EP. Metamers of the ventral stream. *Nat Neurosci.* 2011 Sep;14(9):1195–201.
6. Keshvari S, Rosenholtz R. Pooling of continuous features provides a unifying account of crowding. *J Vis.* 2016 Feb 1;16(3):39–39.
7. Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L. A summary statistic representation in peripheral vision explains visual search. *J Vis.* 2012 Apr 2;12(4):14–14.
8. Gatys L, Ecker AS, Bethge M. Texture Synthesis Using Convolutional Neural Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems 28* [Internet]. Curran Associates, Inc.; 2015 [cited 2017 Oct 18]. p. 262–70. Available from: <http://papers.nips.cc/paper/5633-texture-synthesis-using-convolutional-neural-networks.pdf>
9. Wallis TSA, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *J Vis.* 2017 Oct 1;17(12):5–5.
10. Wallis T, Funke C, Ecker A, Gatys L, Wichmann F, Bethge M. Towards matching peripheral appearance for arbitrary natural images using deep features. *J Vis.* 2017;17(10):786–786.
11. Wilson HR, Cowan JD. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik.* 1973 Sep 1;13(2):55–80.
12. Hermens F, Luksys G, Gerstner W, Herzog MH, Ernst U. Modeling spatial and temporal aspects of visual backward masking. *Psychol Rev.* 2008;115(1):83.
13. Panis S, Hermens F. Time course of spatial contextual interference: Event history analyses of simultaneous masking by nonoverlapping patterns. *J Exp Psychol Hum Percept Perform.* 2014;40(1):129.
14. Clarke AM, Herzog MH, Francis G. Visual crowding illustrates the inadequacy of local vs. global and feedforward vs. feedback distinctions in modeling visual perception. *Front Psychol.* 2014;5.
15. Li Z. Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Netw Comput Neural Syst.* 1999;10(2):187–212.
16. Zhaoping L. V1 mechanisms and some figure–ground and border effects. *J Physiol-Paris.* 2003;97(4):503–15.
17. Cao Y, Grossberg S. A laminar cortical model of stereopsis and 3D surface perception: closure and da Vinci stereopsis. *Spat Vis.* 2005 Nov 1;18(5):515–78.
18. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. 2017;
19. Grossberg S. Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Netw.* 2017;87:38–95.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015 May 28;521(7553):436–44.
21. Lin HW, Tegmark M, Rolnick D. Why Does Deep and Cheap Learning Work So Well? *J Stat Phys.* 2017 Sep 1;168(6):1223–47.
22. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems.* 2012. p. 1097–105.
23. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: <https://www.tensorflow.org/>

24. Chaney W, Fischer J, Whitney D. The hierarchical sparse selection model of visual crowding. *Front Integr Neurosci.* 2014;8.
25. Whitney D, Haberman J, Sweeny TD. From textures to crowds: multiple levels of summary statistical perception. *New Vis Neurosci.* 2014;695–710.
26. Manassi M, Whitney D. Multi-level Crowding and the Paradox of Object Recognition in Clutter. *Curr Biol.* 2018 Feb 5;28(3):R127–33.
27. Fischer J, Whitney D. Object-level visual information gets through the bottleneck of crowding. *J Neurophysiol.* 2011;106(3):1389–98.
28. Nandy AS, Tjan BS. Saccade-confounded image statistics explain visual crowding. *Nat Neurosci.* 2012;15(3):463–9.
29. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis.* 2012 Sep 1;12(10):13–13.
30. 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.
31. Harrison WJ, Bex PJ. A Unifying Model of Orientation Crowding in Peripheral Vision. *Curr Biol.* 2015 Dec 21;25(24):3213–9.
32. Dayan P, Solomon JA. Selective Bayes: Attentional load and crowding. *Vision Res.* 2010 Oct 28;50(22):2248–60.
33. Pachai MV, Doerig AC, Herzog MH. How best to unify crowding? *Curr Biol.* 2016 May 9;26(9):R352–3.
34. Agaoglu MN, Chung ST. Can (should) theories of crowding be unified? *J Vis.* 2016;16(15):10–10.
35. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis.* 2016 Feb 1;16(3):35–35.
36. Manassi M, Hermens F, Francis G, Herzog MH. Release of crowding by pattern completion. *J Vis.* 2015;15(8):16–16.
37. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. *J Vis.* 2013;13(13):10–10.
38. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. *J Vis.* 2012;12(10):13–13.
39. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. *J Vis.* 2016;16(3):35–35.
40. Rosenholtz R, Yu D, Keshvari S. Challenges to pooling models of crowding: Implications for visual mechanisms. *J Vis.* 2019;19(7):15–15.
41. Farzin F, Rivera SM, Whitney D. Holistic crowding of Mooney faces. *J Vis.* 2009;9(6):18–18.
42. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Res.* 1973;13(4):767–82.
43. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. *J Exp Psychol Hum Percept Perform.* 2017;43(4):690.
44. 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.
45. Toet A, Levi DM. The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Res.* 1992;32(7):1349–57.
46. 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.
47. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. *J Vis.* 2009;9(12):13–13.
48. 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;
49. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *European conference on computer vision.* Springer; 2014. p. 818–33.

50. 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.
51. Girshick R, Radosavovic I, Gkioxari G, Dollár P, He K. Detectron. 2018.
52. 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.
53. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
54. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol*. 2014;10(11):e1003915.
55. Kietzmann TC, McClure P, Kriegeskorte N. Deep neural networks in computational neuroscience. *BioRxiv*. 2018;133504.
56. VanRullen R. Perception science in the age of deep neural networks. *Front Psychol*. 2017;8:142.
57. 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.
58. 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.
59. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychol Rev*. 2017;124(4):483.
60. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: *Advances in neural information processing systems*. 2017. p. 3856–66.
61. 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.
62. Mathôt S, Schreij D, Theeuwes J. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behav Res Methods*. 2012;44(2):314–24.

Curriculum vitae

Personal and Professional details

Name: Alban Bornet
Date of Birth: June 30, 1989
Nationality: Swiss

Positions: PhD Student at the Laboratory of Psychophysics
Teaching Assistant in Physics

Institution: Brain Mind Institute
School of Life Sciences
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne - Switzerland

Phone: +41 (0)78 868 40 81
E-mail: alban.bornet@epfl.ch

Education

Sep 2008 - Jul 2011: Bachelor in Life Sciences at EPFL

Sep 2011 - Jul 2012: Bridge year in physics (essentially quantum physics courses to take theoretical physics during the Master)

Sep 2012 - Jul 2014: Master in Neuroscience and Theoretical physics at EPFL

Jul 2014 - Sep 2014: Internship studying waste management in different healthcare centers in Lomé, Togo

Sep 2014 - Jul 2015: Master Project in the Laboratory of Computational Neuroscience, under the supervision of Moritz Deger

Jul 2016 - Feb 2021: Ph.D. in Neuroscience at the Laboratory of Psychophysics (EPFL) under the supervision of Prof. Michael Herzog

Employment history

Sep 2016 - Nov 2019: Worked as a teaching assistant in physics and mathematics during my studies at EPFL

Jul 2016 - Feb 2021: Ph.D. in Neuroscience at the Laboratory of Psychophysics (EPFL) under the supervision of Prof. Michael Herzog

Publications

Publications in peer-reviewed scientific journals

- Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., & Herzog, M. H. (2019). Beyond Bouma's window: How to explain global aspects of crowding? *PLOS Computational Biology*, 15(5), e1006580. Open access link: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006580>
- Doerig, A.[†], Bornet, A.[†], Choung, O.H. and Herzog, M.H., 2020. Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research*, 167, pp.39-45. Open access link: <https://infoscience.epfl.ch/record/275633>
[†]Equal contributions
- Bornet, A., Kaiser, J., Kroner, A., Falotico, E., Ambrosano, A., Cantero, K., Herzog, M.H. and Francis, G., 2019. Running large-scale simulations on the Neurorobotics Platform to understand vision-the case of visual crowding. *Frontiers in neurorobotics*, 13, p.33. Open access link: <https://www.frontiersin.org/articles/10.3389/fnbot.2019.00033/full>

Preprints

- Doerig, A.[†], Bornet, A.[†], Choung, O. H., & Herzog, M. H. (2019). Crowding Reveals Fundamental differences in Local vs. Global Processing in Humans and Machines. *BioRxiv*, 744268. Link: <https://doi.org/10.1101/744268>
[†]Equal contributions

Peer-reviewed conference proceedings

- Choung, O.H. [†], Doerig, A.[†], Bornet, A., & Michael, M.H. (2019). Recurrent Architectures are Needed for Human-like Global Processing. *NeurIPS workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM)*
[†]Equal contributions

Contributions to international conferences

- Using the Neurorobotics Platform to explain global processing in visual crowding. Bornet, A., Kroner, A., Kaiser, J., Scholz, F., Francis, G., Herzog, M.H., European Conference on Visual Perception (ECVP), 2018
- Shrinking Bouma's window: visual crowding in dense displays. Bornet, A., Doerig, A., Van der Burg, E., Herzog, M.H., European Conference on Visual Perception (ECVP), 2019

Posters

- Crowding asymmetries in a neuronal model of image segmentation. Bornet, A., Doerig, A., Herzog, M.H., Francis, G., Vision Sciences Society Meeting (VSS), 2017

- Perceptual grouping and segmentation: uncrowding. Francis, G., Bornet, A., Doerig, A., Herzog, M.H., Vision Sciences Society Meeting (VSS), 2017
- Crowding asymmetries explained by a model of image segmentation. Bornet, A., Doerig, A., Herzog, M.H., Francis, G., European Conference on Visual Perception (ECVP), 2017
- Capsule Networks, but not Convolutional Networks, May Explain Global Configurational Effects. Doerig, A., Bornet, A., Herzog, M. H., EPFL-Google Research Day, 2018
- Capsule networks, but not convolutional networks, explain global configurational visual effects. Doerig, A., Bornet, A., Herzog, M. H., European Conference on Visual Perception (ECVP), 2018

Teaching activities

Teaching assistant:	Analysis III (Sep 2014 - Jun 2015) Physics III & IV (Sep 2016 - Jun 2019)
Student supervision:	Bachelor & Master student supervisor (4 students in total)

Prizes, awards, fellowships

Ingénieurs du Monde fellowship for an internship in Togo (Jun 2014 - Sep 2014)

Personal skills

Digital skills:	Strong abilities in Mathematics, Physics, Machine Learning and Neuroscience Fluent in Python and Matlab. Basic C++ and C Fluent in TensorFlow2, PyTorch Experienced use of modern source control (Git)
Language skills:	French (native), English (fluent), German (basic)
Extra-Professional Activities:	Musician - drummer/singer in three bands