EPFL

Thèse n° 8947

Spatial and temporal integration of visual features

Présentée le 9 décembre 2021

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

pour l'obtention du grade de Docteur ès Sciences

par

Oh-Hyeon CHOUNG

Acceptée sur proposition du jury

Prof. B. D. McCabe, président du jury
Prof. M. Herzog, directeur de thèse
Prof. R. Coen-Cagli , rapporteur
Prof. Y. Yeshurun , rapporteuse
Prof. F. C. Hummel, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

Acknowledgments

I deeply thank my Ph.D. supervisor, Michael H. Herzog, for his full support of my vague and ambitious research plans and for his guidance. Thanks to you, I could learn, explore, and acquire a comprehensive way of thinking and expressing ideas. I enjoyed every single moment of sharing scientific as well as non-scientific discussions with you.

I appreciate the committee members, Prof. McCabe, Prof. Coen-Cagli, Prof. Yeshurun, and Prof. Hummel, for their fruitful comments and discussions. I truly believe your comments and insights improved my work significantly.

Many thanks to my dear LPSY members, without you, I would not be able to walk through this exciting but challenging journey. I would especially like to thank Marc Repnow, for his big supports on technical, scientific, and non-scientific advice. A big Merci to Aline, we started at the same time, shared the entire Ph.D. together. You were my bestie, teacher, pacemaker, supporter, and everything. I have learned a lot from you, and you definitely made my Ph.D. life fruitful and full of love. I have no words to say thank you enough. To Alban, the most brilliant and insightful person I have ever known, a Python master and a philosopher. All the projects I have done with you were full of pleasure. Dhanyawad Harshitha, I know Hindi may not be your favorite language, but still, I do not know how else to thank you enough. You are the smartest and the most motivated researcher and the best ‘storyteller’. Thanks to you, I could learn theories and practices of Adversarial Examples well. Finally, a huge thank you to Janir for your immense support, Adrien for your brilliant insights, and Marc Lauffs for guiding me to the first step of LPSY and research life.

To my strong and “professional” climbers, Edu, Lixia, Inès, Evelyne, Nici, Colin, Arnaud, Wei, Sailan, and all, thanks to you, I could keep both of my physical and mental health intact, especially in this special time of the pandemic.

To my Friday Sat/CH crews, Tobias, Niels, Anhchi, Gizay, Dario, Pol, thanks to you, I could always be relieved. You guys are the best.

To my flatmates Sohyeong, Jihyun, Yeonju, and everyone in the Korean community, especially, Jisoo, Myeongseop, Sukho, SeongRyong, Joowon, Umji, Hwayeong, for your huge emotional supports and all lifelong advice.

To Hyomin, Soorin, Ahyun, and Hyewon, for supporting me from far, for being my best friends for 10+ years, and for everything you have done for me. Without you, I would not be who I am now.

To my lovely sister Jihyeon, you are the wisest person in this universe. Despite you being younger than me, you have always supported me with the right words and gave me the best advice of that moment. I appreciate it a lot, and I assure you that you will get what you deserve.

Last but not least, a special thank you to my parents for everything you have done for me. Thanks to your full support, including that on my spontaneous and uncertain decisions, I could always challenge myself to new things with confidence. Your abundant experiences were always the milestone for my future decision, and your loves were the sanctuary of my mind. I cannot thank you enough, I love you!

Lausanne, 11th November 2021

Abstract

Visual processing can be seen as the integration and segmentation of features. Objects are composed of contours that are integrated into shapes and segmented from other contours. Information also needs to be integrated to solve the ill-posed problems of vision. For example, in the “color” perception of an object, illuminance needs to be discounted, requiring large-scale integration of luminance values. Whereas there is little controversy about the crucial role of integration, very little is known about how it really works. In this thesis, I focused on large-scale spatiotemporal information using two paradigms. First, I used the Ternus-Pikler display (TPD) to understand non-retinotopic, temporal integration, and then I used crowding to understand spatial integration across, more or less, the entire visual field.

Motions of object parts are perceived relative to the specific object. For example, a reflector on a bicycle wheel seems to rotate even though it is cycloidal in retinotopic coordinates. This is because the reflector’s motion is subtracted from the bike’s horizontal motion. Instead of bike motion, I used the TPD, which is perfectly suited to understand non-retinotopic processing. There are two possibilities of how information may be integrated non-retinotopically: either based on attentional tracking, e.g., of the reflector’s motion, or relying on inbuilt automated mechanisms. I showed that attentional tracking does not play a major role for non-retinotopic processing in the TPD. Second, I showed that invisible retinotopic information can strongly modulate the visible, non-retinotopic percept, further supporting automated integration processes.

Crowding occurs when the perception of a target deteriorates because of the surrounding elements. It is the standard situation in everyday vision, since elements are rarely encountered in isolation. The classic model of vision integrates information from low-level to high-level feature detectors. By adding flankers, this model can only predict performance deterioration. However, this prediction was proven wrong because flankers far from the target can even lead to a release of crowding. Integration across the entire visual field is crucial. Here, I systematically investigated the characteristics of this large-scale integration. First, I dissected complex multi-flanker configurations and showed that low-level aspects play only a minor role. Configural aspects and the Gestalt principle of Prägnanz seem to be involved instead. However, as I showed secondly, the basic Gestalt principles fail to explain our results. Lastly, I tested several computational models, including *one-stage feedforward models* that integrate information within a local area or across the whole visual field, and *two-stage recursive models* that integrate global information and then explicitly segment elements. I showed that all models fail, unless they take explicit grouping and segmentation processing into accounts, such as capsule networks and the Laminart model.

Overall, spatial and temporal integration is rather a complex inbuilt automated mechanism, and integration occurs across the whole visual field, contrary to most classic and recent models in vision. Moreover, global integration can only be reproduced by two-stage models, which process grouping and segmentation. To better understand perception, we need to consider models that group elements by multiple processes and recursively segment other groups explicitly.

Keywords

Human vision, Perceptual organization, Non-retinotopic processing, Crowding, Model of vision, Grouping, Segmentation, Feedforward networks, Convolutional Neural Networks, Recurrent models, Ternus-Pikler display, Vernier stimulus

Résumé

La vision humaine intègre l'information visuelle pour identifier et segmenter les objets dans le champ visuel. Un rôle prépondérant de cette intégration est de résoudre les problèmes inhérents à la vision qui ne peuvent pas être résolus par les yeux uniquement. Par exemple, pour percevoir la couleur d'un objet, le cortex visuel doit prendre en compte l'illumination ambiante de cet objet, ce qui nécessite une intégration contextuelle des valeurs de luminance dans la totalité du champ visuel. Alors que le rôle crucial de cette intégration de l'information est peu controversé dans les modèles de la vision humaine, on sait encore très peu de choses sur son fonctionnement réel. Dans cette thèse, je me suis concentrée sur comment la vision humaine intègre l'information sur de grandes échelles d'espace et de temps, en utilisant deux paradigmes. Tout d'abord, j'ai utilisé le paradigme de Ternus-Pikler (« Ternus-Pikler Display » ; TPD) pour comprendre l'intégration temporelle de l'information visuelle lorsqu'elle se déroule de manière non-rétinotopique. Puis, j'ai utilisé l'encombrement visuel (« visual crowding ») pour comprendre comment et pourquoi l'information visuelle peut être intégrée sur des distances qui couvrent presque l'ensemble du champ visuel.

Les mouvements des parties d'un objet sont perçus par rapport à l'objet en question. Par exemple, un catadioptre sur une roue de bicyclette qui avance semble se mouvoir en cercles, alors que, considérant les coordonnées rétinotopiques, ce mouvement trace une cycloïde. Cela est dû au fait que la vision humaine soustrait le mouvement horizontal du vélo à celui du catadioptre. En lieu et place du mouvement du vélo, j'ai utilisé le TPD, un paradigme qui est parfaitement adapté pour comprendre le traitement non-rétinotopique de l'information visuelle. Il existe deux possibilités: soit l'information est intégrée sur la base d'un suivi volontaire, ou « attentionnel », par exemple dirigé sur le mouvement du catadioptre ; soit elle est intégrée sur la base de mécanismes automatiques qui sont déclenchés directement par l'information visuelle. J'ai montré que, dans le TPD, le suivi attentionnel ne joue pas un rôle majeur dans le traitement non-rétinotopique de l'information visuelle. De plus, j'ai montré que l'assimilation d'information rétinotopique invisible module fortement la perception non-rétinotopique de ce qui est visible ; ce qui renforce la théorie de processus d'intégration automatisés.

L'encombrement visuel (« visual crowding ») se produit lorsque la perception d'un objet-cible se détériore en présence d'autres éléments avoisinants (les « distracteurs »). Il s'agit d'une situation standard de la vision quotidienne, puisque les objets sont rarement perçus de manière isolée. Les modèles classiques considèrent que la vision humaine décompose l'information visuelle en éléments basiques puis combine ces éléments localement pour détecter des formes et objets de plus en plus complexes. Dans ces modèles, l'information va dans un seul sens (« feedforward » ; des contours basiques aux objets, en passant par des formes de plus en plus complexes). Ce type de modèles prédit que, dans l'encombrement visuel, ajouter des distracteurs ne peut que détériorer la perception de l'objet-cible. Cependant, des expériences montrent qu'il est possible d'aligner des distracteurs autour de la cible sur de grandes distances et obtenir, au contraire, une amélioration de la performance. Ici, contrairement à ce que les modèles classiques prédiraient, l'intégration de l'information dans l'ensemble du champ visuel jour un rôle crucial dans la perception visuelle. Pour comprendre pourquoi, j'ai étudié de manière systématique les caractéristiques de cette intégration à longue distance. Tout d'abord, j'ai utilisé des configurations de distracteurs complexes dans des expériences d’encombrement visuel. En modifiant méticuleusement ces configurations, j’ai montré que leurs caractéristiques locales ne jouent qu'un rôle mineur. En revanche, les aspects globaux de ces configurations et la loi Gestalt de Prägnanz semblent jouer un rôle majeur (alors que les lois Gestalt basiques n’expliquent pas les résultats). Enfin, j'ai testé plusieurs modèles de la vision humaine, comprenant des modèles classiques, qui n’intègrent l’information que dans un seul sens et localement, un modèle qui intègre l’information de manière globale et dans un seul sens et, finalement, des modèles qui intègrent l’informations de manière globales et récurrente, et qui sont capables de segmenter les éléments visuels en différents groupes perceptuels. J'ai montré que les seuls modèles capables d’expliquer mes résultats expérimentaux sont ceux qui contiennent des processus explicites de groupement visuel et de segmentation, comme par exemple les réseaux de capsules (« capsule networks ») ou le modèle « Laminart ».

Dans l'ensemble, ma thèse montre que l'intégration spatiale et temporelle de l’information visuelle est un mécanisme complexe et automatisé qui se produit dans la totalité du champ visuel, contrairement à ce que prédisent la plupart des modèles classiques, et même les plus récents, de la vision humaine. De plus, mes résultats expérimentaux ne peuvent être reproduits que par des modèles dans lesquels des processus de groupement visuel et de segmentation influencent de manière récurrente comment et quelle information est intégrée. Pour mieux comprendre la vision humaine, nous devons envisager des modèles qui incluent ce genre de mécanismes.

Mots-clés

Vision humaine, Organisation perceptive, Traitement non-rétinotopique, Encombrement visuel, Modèle de la vision, Groupement visuel, Segmentation, Réseaux feedforward, Réseau neuronal de convolution, Modèles récurrents, Paradigme de Ternus-Pikler, Stimulus de Vernier

Preface

In this thesis, I present the results of a series of projects conducted during my doctoral studies. My projects are focused on the spatial and temporal integration of visual features. The list of studies is included below and annotated with my personal contributions. At the time of writing, four studies are published (or accepted), and two are in preparation. Manuscripts of published articles and preprints of the submitted manuscripts (or manuscripts in preparation) are provided in the main body of the thesis with the permission of the copyright holders. In addition, one submitted and two published projects on different topics are presented in the Appendix.

Included in thesis main body :

1. Lauffs, M. M.*, **Choung, O. H.***, Öğmen, H., Herzog, M. H., & Kerzel, D. (2019). Reference-frames in vision: contributions of attentional tracking to nonretinotopic perception in the Ternus-Pikler display. *Journal of vision*, *19*(12):7, 1-15.
   I designed, coded, and conducted the experiment and analyzed the data with L.M.M., and wrote the manuscript together with all the authors.

2. Lauffs, M. M., **Choung, O. H.**, Öğmen, H., & Herzog, M. H. (2018). Unconscious retinotopic motion processing affects non-retinotopic motion perception. *Consciousness and cognition*, *62*, 135-147.
   I collected and analyzed the data with L.M.M, designed and coded the computational model, and wrote the manuscript together with all the authors.

3. 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.
   I designed Experiment 2, and coded and ran the shape-biased ResNet50, and wrote the manuscript together with all the authors.

4. Ruthemann, N.*, **Choung O. H.***, & Herzog, M. H., (*in prep*). What crowds in crowding?
   I designed and coded the experiment, analyzed the data with N.R., and wrote the manuscript together with all the authors.

5. **Choung, O. H.,** Bornet, A., Doerig, A., & Herzog, M. H., (2021). Dissecting (un)crowding, *Journal of Vision*, 21(10):10, 1-20.
   I designed, coded, conducted the experiment, analyzed the data, coded and ran model simulations with A.B., and wrote the manuscript with M.H.H.

6. **Choung, O. H.**, Rashal, E., Kunchulia, M., & Herzog, M. H. (*in prep*). Basic gestalt principles cannot explain Uncrowding.
   I designed the experiment with ER, coded the experiment, collected the data with M.K., analyzed the data, coded and ran the model simulation, and wrote the manuscript with M.H.H.

Contributions in other fields :

1. Bornet, A., **Choung, O. H.**, Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (2021). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Jornal of vision*, 21(12):10, 1-25. (Appendix A)
   I collected and analyzed the data with A.B., and wrote the manuscript together with all the authors.

2. Machiraju, H.*, **Choung, O. H.***, Frossard, P., & Herzog, M. H. (2021). Bio-inspired Robustness: A Review. *arXiv preprint arXiv:2103.09265*. (Appendix B)
   H.M. and I planned and reviewed papers, coded and ran the attacks, and wrote the manuscript with P.F. and M.H.H.

3. Cretenoud, A. F., Barakat, A., Milliet, A., **Choung, O. H.**, Bertamini, M., Constantin, C., & Herzog, M. H. (2021). How do visual skills relate to action video game performance?. *Journal of vision*, *21*(7): 10, 1-21. (Appendix C)
   I planned and coded the Crowding experiment with A.F.C. and A.B., and revised the manuscript with all the authors.

* denotes equal contributions.

Contents

Acknowledgments .... **v**

Abstract .... **vi**

Keywords .... **vii**

Résumé .... **viii**

Mots-clés .... **ix**

Prefacex

Contents **12**

Chapter 1 Introduction **13**

- 1.1 The classic model of vision: feedforward and hierarchical visual processing 15
- 1.2 Retinotopic processing vs. Non-retinotopic processing 16
- 1.3 Visual crowding vs. Uncrowding 18
- 1.4 Modeling approaches 21
- 1.5 Aims of the present work 23

Chapter 2 Non-retinotopic perception **25**

- 2.1 Non-retinotopic motion perception is not merely an attentional tracking 26
- 2.2 Unconscious retinotopic information affects non-retinotopic motion perception 47

Chapter 3 Visual crowding **69**

- 3.1 Feedforward Convolutional neural networks cannot explain (un)crowding 70
- 3.2 How does target information integrate in crowding? 84
- 3.3 Dissecting (un)crowding 103
- 3.4 Basic Gestalt principles cannot explain (un)crowding 136

Chapter 4 Conclusions **158**

- 4.1 Summary 158
- 4.2 Future development 163

References **164**

Appendix **174**

Chapter 1 Introduction

We perceive the world effortlessly, but vision is extremely complex. About one-third of the cerebral cortex is dedicated to this complex process (Van Essen et al., 2001). Everyday vision is a consequence of grouping and segmentation that serve as a new perception unit (Wagemans, Elder, et al., 2012). Visual elements are first integrated into objects, and then parts of the objects are perceived based on the object they belong to. Fig. 1.1.1a shows a mesh fence in front of a house. On the retina, the contours of the mesh fence and the house overlap. When processed in a local and feedforward fashion, nearby elements from both the mesh fence and the house will be pooled together and intermixed, making it impossible to segregate them. Instead, a global and recursive process that can determine which elements belong together is needed.

As mentioned, objects, rather than low-level visual features, are the basis of human visual perception. Here are two other examples. A classic example of perceptual grouping phenomena is the hidden Dalmatian picture by photographer R. C. James (Fig. 1.1.1c). At first glance, the dots on the left image seem to be randomly distributed. However, as soon as we notice the Dalmatian, as in the right panel (Fig. 1.1.1c), it is impossible to ignore the Dalmatian. Other grouping cues, such as coherent motion, can also induce perceptual grouping. In this case, dots moving together – in coherence – can induce the percept of a running Dalmatian, making the individual dots hardly distinguishable. Perceptual grouping can also be induced by proximity or coherency (Todorović, 2007). For example, in Fig. 1.1.1b, lines are automatically integrated to be perceived as sneakers. Yet, the lines within the same group have features beyond proximity or coherency. The examples are hardly explained by combinations of its parts, as in Aristotle's (B.C. 384-322) famous adage 'The whole is more than the sum of its parts'.

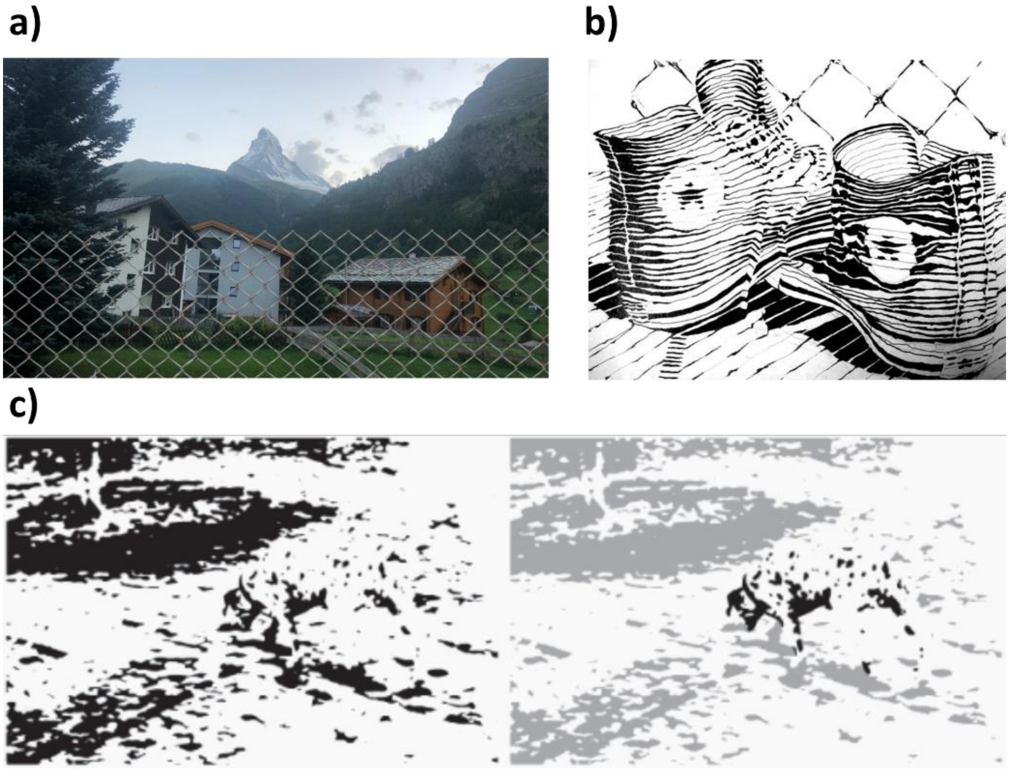


Figure 1.1.1. Grouping and segmentation are necessary and automatic. a) In daily visual situations, visual elements are grouped, and each group serves as the fundamental perceptual unit for visual processing. Overlapped contours can be segregated depending on their group membership. For example, a mesh fence in front of a house leads to overlapped contours. However, we perceive a fence and a house in different perceptual groups, despite the intermingled local features. This process is automatic and hard to avoid. b) A pair of sneakers drawn only with strokes. c) A hidden Dalmatian picture by photographer R. C. James (left), and the contrast increased picture (right).

Perception depends on grouping and segmentation. The important question here is when and how we integrate low-level features to objects. In this thesis, I used two psychophysical paradigms, namely non-retinotopic processing and crowding, to investigate 1) the characteristics of object-based perception, and 2) how features integrate.

1.1 The classic model of vision: feedforward and hierarchical visual processing

In the classic model of vision (Fig. 1.1.2), visual features are processed in a local and feedforward manner (Hubel et al., 1978; Hubel & Wiesel, 1959, 1962; Riesenhuber & Poggio, 1999). The retina receives visual inputs, which are then propagated to the Lateral Geniculate Nucleus (LGN), and to the visual cortex. Low-level features, such as orientation, are processed in the primary visual cortex (V1). Then, edge and line information is pooled and proceeded to V2 to detect corners and simple shapes. Finally, objects are detected in the inferior temporal cortex (IT; Kravitz et al., 2013). Object recognition can be very fast (under 150ms) in both human (Thorpe et al., 1996) and monkey experiments (Thorpe et al., 2001). Such a rapid process favors feedforward processing (Fabre-Thorpe, 2011). However, this assumption has been challenged (Pollen, 1999), and theories are proposed that visual processing is not only feedforward but recursive (Lamme & Roelfsema, 2000).

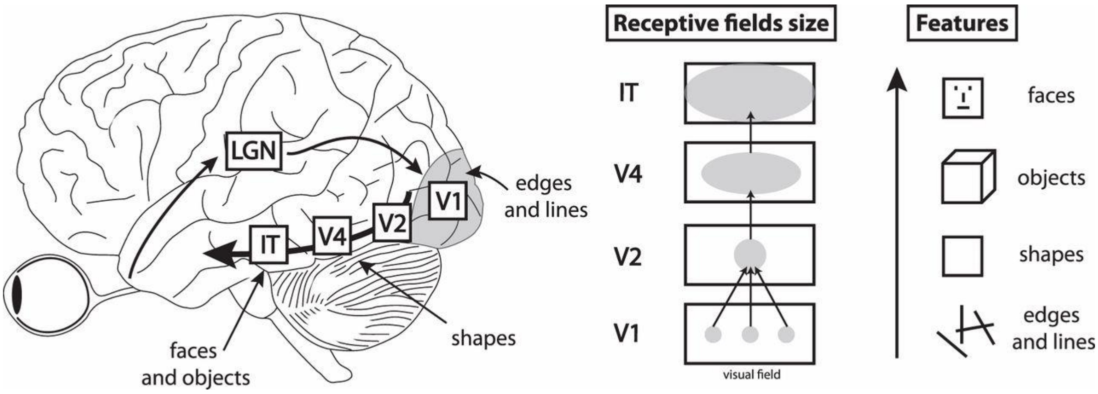


Figure 1.1.2. The classic model of vison processes visual stimuli in a series of visual areas, from low-level to high-level, in a local, feedforward, and hierarchical manner. First, the retina receives inputs and proceeds signals to the visual cortex. The early visual areas are sensitive to low-level features. The primary visual cortex (V1) is sensitive to low-level features such as edges and lines. Then, higher visual areas, with larger receptive fields, are sensitive to more complex features, such as objects. Higher levels of processing are determined by low-level processing. Reprinted from Manassi et al. (2013)

1.2 Retinotopic processing vs. Non-retinotopic processing

The first stages of the visual system are *retinotopic* (Fig. 1.2.1). At the entrance of the visual system, the retina encodes the 3D outer world in a 2D upside-down image. Then the visual information proceeds to the cortex through LGN and V1. In this stage, the inputs reflect retinal geometry, such that the relative distances between different points are reserved (Tootell et al., 1998). For example, in Fig. 1.2.1, P1 and P2 are encoded closer than P1 and P3 in V1.

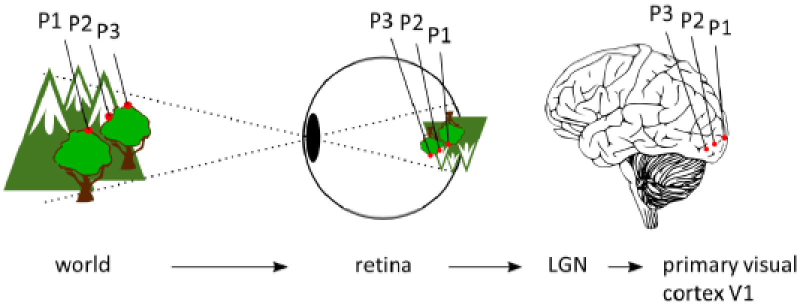


Figure 1.2.1. Early visual processing is retinotopic. In the retina and early visual areas, neighboring positions in the real world are encoded next to each other, as shown by the points P1, P2, and P3. Therefore, relative neighborhood relationships and distances are preserved. Reprinted from Lauffs (2017).

However, perception is usually *non-retinotopic*. The parts of a moving object are perceived relative to the object, rather than in retinotopic coordinates (Agaoglu et al., 2016; Clarke et al., 2016; Duncker, 1929; Johansson, 1950, 1973; Öğmen & Herzog, 2010). For example, the reflector on a bicycle wheel is perceived to rotate, although its motion is cycloidal in retinotopic coordinates (Fig. 1.2.2; Duncker, 1929; Johansson, 1950). The bicycle's horizontal motion is perceptually discounted from the cycloidal retinotopic motion of the reflector. Only the circular non-retinotopic reflector motion is perceived consciously, whereas the cycloidal retinotopic motion is invisible. Similar to ambiguous figures, all elements are visible, only one interpretation is suppressed (Herzog, Hermens, & Öğmen, 2014). Obviously, the percept cannot be explained exclusively by retinotopic inhibition because it depends on non-retinotopic information.

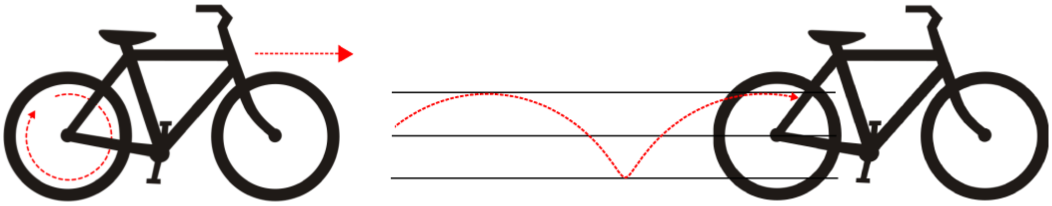


Figure 1.2.2. Reflectors on a bicycle wheel are perceived to rotate (left). However, in retinotopic coordinates, its motion is cycloidal (right). This is due to the reflector motion being perceived relative to the bicycle, so the bicycle works as the reference frame of the reflector motion as suggested by Clarke et al., (2016); Ogmen & Herzog, (2010). Reprinted from Lauffs et al. (2017).

I used a modified version of the TPD to access non-retinotopic perception (Fig. 1.2.3; Boi et al., 2009; Lauffs et al., 2017; Pikler, 1917; Ternus, 1926). In the TPD, two disks are continuously flickering in the center of a screen, and a third disk is added alternatively on the left or right of the other two disks. When the inter-stimulus interval (ISI) is very brief (e.g., 0ms), the third disk appears to jump from the left to the right of the two stationary flickering disks (element motion). When the ISI is larger (e.g., 200ms), the three disks form a perceptual group, and all three disks appear to shift left and right in tandem (group motion). The stimulus positions are the same in screen-based retinotopic coordinates when staring at a central fixation point. When dots are added to the disks, as in Fig. 1.2.3, disks serve as a reference frame for perceiving dot motion (similar to the bicycle works as a reference frame for perceiving the reflector). In element motion conditions (Fig. 1.2.3 left panel), two stationary flickering disks work as the reference frame of the dot motion. Hence, the dots in disks appear to move up-and-down and left-and-right, respectively. The dot percept is *retinotopic*, as the disk center is aligned with the retinotopic coordinate. However, in group motion conditions (Fig. 1.2.3 right panel), three disks moving left-and-right in tandem works as the reference frame. Thus, the dot in the middle disk is perceived to rotate. This dot rotation percept is *non-retinotopic*, because it can only be perceived in disk-centered coordinates; in retinotopic coordinates, the dots still move linearly up-down and left-right.

A two-stage model was suggested (Clarke et al., 2016; Ogmen & Herzog, 2010). In the first stage, disks and dots are detected, and their retinotopic motion vectors are computed. Then, objects are grouped based on the coherent motion and proximity. Each group then produces a common motion vector that serves as a reference frame (e.g., disks' element or group motion). In the second stage, stimuli in retinotopic representations are re-mapped to non-retinotopic representations. For example, the dot motion percept is made by subtracting the disk motion vector from each dot's retinotopic motion vector. This re-mapping step made linear up-down and left-right retinotopic dot motion to be perceived as circular non-retinotopic dot motion. Therefore, non-retinotopic processing requires a global grouping process, which the classic model of vision cannot explain.

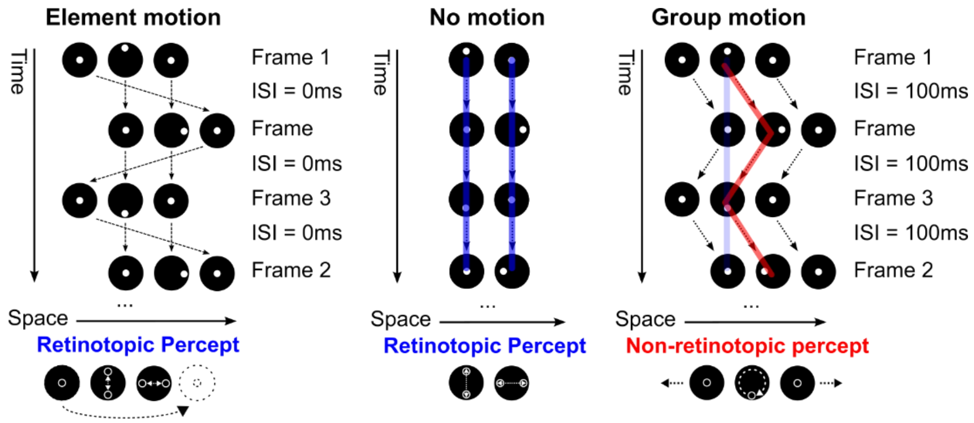


Figure 1.2.3 Ternus-Pikler Display (TPD) with linear retinotopic and circular non-retinotopic dot motion, reprinted from Lauffs et al., (2018). Left: When three disks are presented with a short ISI (< 33ms), element motion is perceived. Participants perceive two flashing disks and one left-right jumping disk (retinotopic percept). Middle: When two disks are presented, no disk motion is perceived, but flickering disks. The dot in the left disk is perceived to move up-and-down and left-and-right in the right disk (blue, retinotopic percept). Right: When three disks are presented with a long ISI (≥ 33ms, e.g. 100ms), the three disks are perceived to move left-and-right in concert ("group motion"), and the dot in the middle disk is perceived to rotate (red, non-retinotopic percept). In the left and right disks the dots are always in the center.

1.3 Visual crowding vs. Uncrowding

In the classic model of vision, perception of a target strongly deteriorates when embedded in context, which is referred to as *crowding* (Fig. 1.3.1, reviews: Herzog et al., 2016; Levi, 2008; Pelli & Tillman, 2008; Strasburger, 2020). For example, in Fig. 1.3.1, while keeping the eyes on the center, the child on the left is easier to perceive than the child on the right. This is because the same-colored signposts next to the child act as flankers, distracting target perception. Crowding is the standard situation in everyday vision because elements are rarely encountered in isolation. Crowding is stronger when the target and flankers share similar features, such as the same contrast polarity (Kooi et al., 1994), color (Kennedy & Whitaker, 2010; Põder, 2007; van den Berg et al., 2007), orientation (Andriessen & Bouma, 1976; Parkes et al., 2001; Wilkinson et al., 1997), motion (Bex & Dakin, 2005; Gheri et al., 2007), spatial frequency (Chung et al., 2001; Põder & Wagemans, 2007), etc. It is often argued that only flankers within a certain spatial window (Bouma's window; ½ of the eccentricity) around the target deteriorate performance (Bouma, 1970; Bouma, 1973; Levi, 2008; Strasburger et al., 1991; Weymouth, 1958). Crowding also has specific characteristics tied to low-level features. For example, flankers in the radial orientation interfere stronger than flankers in the tangential orientation (radial-tangential anisotropy; Chung, 2013; Greenwood et al., 2017; Kwon et al., 2014; Malania et al., 2020; Toet & Levi, 1992), which is explained by the elliptic shape of receptive fields in early visual areas (Hubel et al., 1978; Silson et al., 2018; Toet & Levi, 1992) or by an uneven sampling density in the early visual cortex (Kwon & Liu, 2019; Motter & Simoni, 2007).

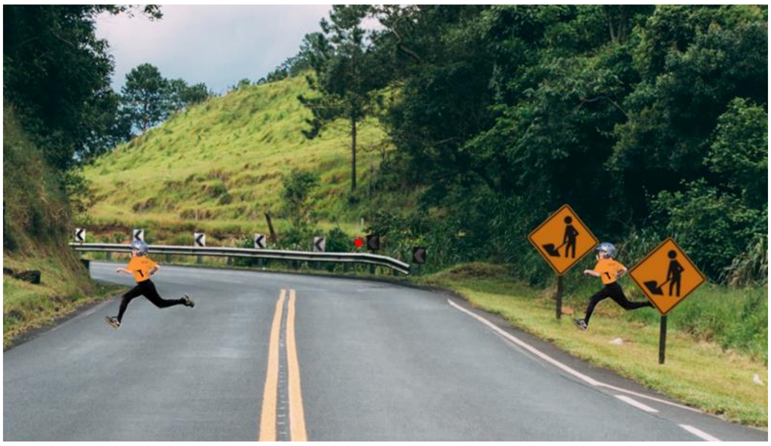


Figure 1.3.1. Classic crowding example. In crowding, target perception deteriorates in the presence of nearby visual elements. When fixating on the central red dot, the child on the left is easy to perceive. However, the child on the right is hardly perceived because of the nearby signposts. Figure reproduced from Doerig, Bornet, et al. (2019)

Based on the above-mentioned characteristics, crowding is traditionally explained by local, feature-specific interactions between the neural representations of the target and its direct neighbors. For example, neurons sharing the same orientation may interact with each other through lateral inhibition, feedforward pooling (e.g., Dakin et al., 2010; Greenwood et al., 2009, 2017; Parkes et al., 2001; Pelli, 2008; Rosenholtz et al., 2012; Solomon et al., 2004), features may be substituted (Huckauf & Heller, 2002; Strasburger et al., 1991;

Strasburger, 2005), or crowding may be mediated by a top-down process (S. He et al., 1996; Montaser-Kouhsari & Rajimehr, 2005; Tripathy & Cavanagh, 2002; Yeshurun & Rashal, 2010). In all these explanations, target information is irretrievably lost at the early stages of visual processing. Thus, crowding research is very much in the spirit of an atomistic view of visual processing, where basic, local processing precedes more complex processing.

However, these explanations break down when the target is presented with complex, instead of simple flanker configurations (e.g., Livne & Sagi, 2007; Manassi et al., 2016; Põder, 2007; Saarela et al., 2009; Sayim et al., 2010; Yeotikar et al., 2011). For example, when a Vernier is presented, it is easy to discriminate the offset direction (left or right). Vernier offset discrimination drops drastically when the vernier is presented within a square (crowding). However, performance comes back almost to the unflanked, Vernier alone condition's level when more squares are added on both sides of the Vernier, which is referred to as *uncrowding.* The Vernier information is recovered likely because, the target and the flankers (i.e., the squares) are segmented in different perceptual groups, which is not the case when only one square is presented (Fig. 1.3.2 left panel). Manassi and colleagues (2013) have shown that the target is easily perceived when the central square is crowded by the other squares.

Similar effects have been shown previously with various stimuli such as Verniers (Manassi et al., 2012, 2013, 2015, 2016; Sayim et al., 2010), Gabors (Levi & Carney, 2009; Livne & Sagi, 2007; Maus et al., 2011; Saarela et al., 2009; Saarela & Herzog, 2008), shapes (Kimchi & Pirkner, 2015), letters (Reuther & Chakravarthi, 2014; Saarela et al., 2010), textures (Herrera-Esposito et al., 2020), as well as in haptics (Overvliet & Sayim, 2016) and audition (Oberfeld & Stahn, 2012). Feature similarity is important but not decisive, since strong crowding also occurs with flankers of different contrast polarity and color (Manassi et al., 2012; Sayim et al., 2008). What matters is the configuration (Livne & Sagi, 2007) of potentially all elements across large parts of the visual field (Herzog et al., 2016; Herzog & Manassi, 2015).

Several models have been proposed for crowing, however, most fail to explain the phenomenon of uncrowding. Doerig and colleagues (2019) compared multiple models of crowding, and showed that only models based on grouping and segmentation processes could explain uncrowding (see Chapter 3.3).

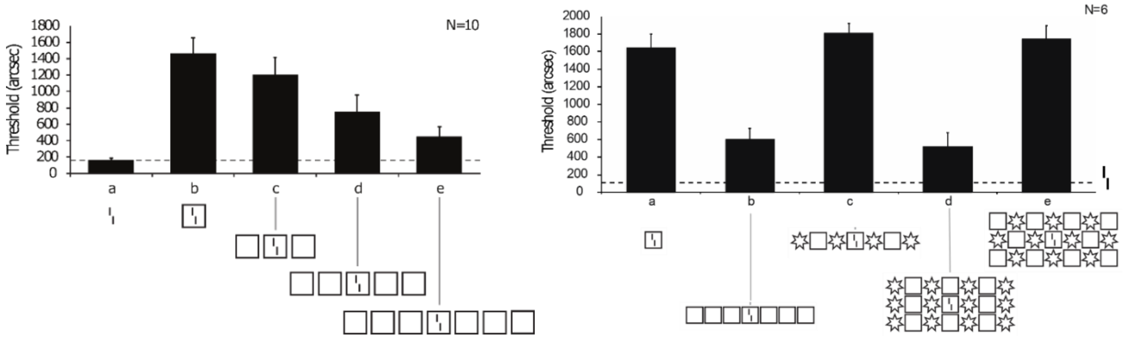


Figure 1.3.2. Uncrowding, left reprinted from Manassi et al. (2013) and right reprinted from Manassi et al. (2016). The y-axis represents Vernier offset thresholds (arcsec), hence, larger values indicate worse performance levels. Left: One square flanker strongly deteriorates Vernier discrimination (classic crowding). When more squares are added, performance improves; that is, the crowding effect is diminished (uncrowding). Right: Uncrowding strongly depends on the configuration. For example, condition d shows a configuration of flankers with a strong uncrowding effect. In comparison, condition e has the same flankers but in a different configuration producing strong crowding.

1.4 Modeling approaches

Therefore, the classic model of vision cannot explain visual perception well. However, the state-of-the-art vision models for both computer vision and human vision follow the classic model of vision, which only vaguely mimic biological vision; Convolutional Neural Networks (CNN; e.g., He et al., 2016; Hinton, 1981; Krizhevsky et al., 2012; LeCun et al., 1988). Feedforward CNNs process visual features based on feedforward and hierarchical local interactions, but CNNs could still attain ‘superhuman performance’ in miscellaneous tasks, such as classification (e.g., He et al., 2016; Krizhevsky et al., 2012), segmentation (Girshick et al., 2018; review: Hafiz & Bhat, 2020), and even video games (e.g., Schrittwieser et al., 2020; Silver et al., 2017). These ffCNNs and the human visual system share several similarities. For example, after training on complex visual datasets such as ImageNet (Deng et al., 2009), ffCNN neural activities show high correlations with human and non-human primate neural activities (Khaligh-Razavi & Kriegeskorte, 2014; Nayebi et al., 2018; Yamins et al., 2014) 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 (Lindsey et al., 2019; Zeiler & Fergus, 2014). Because of these similarities, ffCNNs trained on complex visual tasks were proposed as models of the human visual system (Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann, McClure, & Kriegeskorte, 2018; Nayebi et al., 2018; VanRullen, 2017; Yamins et al., 2014). However, there are dissimilarities. CNNs, unlike humans, are prone to a small amount of noise optimized to cause misclassification, e.g., a single pixel can change the classification (adversarial examples; Szegedy et al., 2014). Importantly, several studies have shown that feedforward CNNs usually rely on local features while humans strongly rely on global shape information (Baker, Lu, Erlikhman, & Kellman, 2018; Brendel & Bethge, 2019; Doerig, Bornet, et al., 2019; Kim, Bair, & Pasupathy, 2019). The bias in CNNs may come from training, as Gerihos and colleagues (2018) suggested. However, the bias can also come from the fundamental structural differences between human and machine vision (see chapter 3.1; Doerig, Bornet, et al., 2020). That is, the feedforward and the local processes cannot fully account for the global shape information. This may suggest a global and recursive model is needed for the human-like object-based computations.

Several recursive grouping models were proposed to address the limitations of feedforward models. For example, Francis and colleagues (2017) suggested that the Laminart model can produce a perceptual group by collinearity grouping, similar to 'association fields’ (Field et al., 1993). The grouping layer in the Laminart model collects input from neurons with similar orientation preferences that are arranged along the preferred direction. And this connects the elements within a local neighborhood to make a perceptual group. Also, Linsley, Kim, and colleagues (Kim et al., 2019; Linsley et al., 2020) suggested that lateral (horizontal) recursive connections can support low-level feature-based incremental grouping. On the other hand, top-down connections rescue learning on tasks with higher-level object information, such as alphabets. Finally, Vacher and colleagues (Vacher et al., 2020, 2019; Vacher & Coen-Cagli, 2019) suggested that flexible pooling of neurons, such as in divisive normalization (Coen-Cagli et al., 2015), may contribute to grouping.

1.5 Aims of the present work

Daily vision is a consequence of grouping and segmentation. The percept is made in two steps: first, elements (e.g., reflector, part of mesh fence) are integrated into an object (e.g., bicycle, fence, or house), then, objects are segmented so that the parts are perceived based on their own membership. It is crucial to understand how elements are integrated across space and time. A feedforward and hierarchical traditional vision model may explain the spatial and temporal integration perfectly. It is questionable because, as in the fence and the house example, a recursive signal is needed to provide information about each visual element's object membership. Therefore, this thesis aims to, 1) observe the properties of object-based perception both in space and time by using well-controlled psychophysical experiments; 2) test whether the state-of-the-art CNN models can magically enable the grouping and segmentation despite their feedforward hierarchical nature. I used two paradigms, namely non-retinotopic perception and visual crowding, perfectly-suited processes to understand spatial and temporal integrations.

In the following sections, I examine the properties of spatiotemporal integration using non-retinotopic perception (chapter 2) and spatial integration using visual crowding (chapter 3).

Chapter 2 Non-retinotopic perception

In this chapter, I studied how spatiotemporal integration affects motion perception using the Ternus-Pikler display (TPD). In the first study, I showed that non-retinotopic motion perception is not merely attentional tracking. In the second study, I showed that unconscious retinotopic information can survive and affect conscious non-retinotopic motion perception.

2.1 Non-retinotopic motion perception is not merely an attentional tracking

Full citation: Lauffs, M. M.*, **Choung, O. H.***, Öğmen, H., Herzog, M. H., & Kerzel, D. (2019). Reference-frames in vision: contributions of attentional tracking to nonretinotopic perception in the Ternus-Pikler display. *Journal of vision*, *19*(12):7, 1-15.

Summary:

Dot motion perception in the TPD can change strongly depending on its reference frame. In TPD, it seems that the group motion of the disks serves as a reference frame for the motion processing of the white dots (Figs. 1.2.3). However, it could well be that dot motion occurs because of attentional tracking.

In this work, we showed that non-retinotopic motion perception was not merely attentional tracking. We used the modified TPD and highlighted the relevant disk in distinct colors (e.g., black vs. turquoise) or a distinct shape (e.g., square vs. circle). The results show that, despite improved performance with the distinct disk color, performance was still lower than expected from tracking. This suggests that attentional tracking cannot fully explain non-retinotopic processing. Moreover, when the target disk had a distinct feature (frame color difference, shape difference, etc.), performance improved compared to the no cue condition. However, performance was much lower than in the three disk condition (group motion condition), which is expected from a tracking mechanism. Interestingly, when only a single disk was shown (1-disk condition), performance was worse than with group motion, even though the target was well tracable. These results strongly suggest that the non-retinotopic integration is not based on attentional tracking.

Postprint of the article published in *Jornal of Vision*:

Reference-frames in vision: Contributions of attentional tracking to non-retinotopic perception in the Ternus-Pikler display

Marc M. Lauffs[a†], Oh-Hyeon Choung[a†*], Haluk Öğmen[c], Michael H. Herzog[a] & Dirk Kerzel[b]

[a] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, http://lpsy.epfl.ch, marc.lauffs@gmail.com, oh-hyeon.choung@epfl.ch, michael.herzog@epfl.ch

[b] Faculté de Psychologie et des Sciences de l'Éducation, Université de Genève, Geneva, Switzerland, dirk.kerzel@unige.ch

[c] Department of Electrical and Computer Engineering, University of Denver, CO, USA, haluk.ogmen@du.edu

†Both authors contributed equally to this manuscript. *Corresponding author

Abstract

Perception depends on references frames. For example, the "true" cycloidal motion trajectory of a reflector on a bike's wheel is invisible because we perceive the reflector motion relative to the bike's motion trajectory, which serves as a reference frame. To understand such an object-based motion perception, we suggested a "two-stage" model in which first reference frames are computed based on perceptual grouping (bike) and then features are attributed (reflector motion) based on group membership. The overarching goal of this study was to investigate how multiple features (i.e. motion, shape and color) interact with attention to determine retinotopic or non-retinotopic reference frames. We found that, whereas tracking by focal attention can generate non-retinotopic reference-frames, the effect is rather small compared to motion-based grouping. Taken together, our results support the two-stage model and clarify how various features and cues can work in conjunction or in competition to determine prevailing groups. These groups in turn establish reference-frames according to which features are processed and bound together.

Keywords

Non-retinotopic processing; reference frame; attentional tracking; motion perception

Introduction

According to behavioristic theories, stimuli generate *sensations*, which are stored in memory. Correlations between stimuli (classical conditioning) or stimuli and behavioral responses (operant conditioning) cause these sensations to be *associated* with each other to generate complex representations. However, many predictions of this approach failed when subjects responded in disagreement with the "association strengths" of stimuli. Several theoreticians then resorted to the concept of *attention* as pointed out by Koffka (1922, p. 535): *"wherever there is an effect that cannot be explained by sensation or association, there attention appears upon the stage"*. A key shortcoming of behavioristic theories is their inability to define the stimulus independent of the observer, a problem known as the "stimulus definition" problem. To deal with this problem, Tolman introduced "intervening variables" to account for how internal states of the observer can modify the stimulus (Tolman, 1938). In contrast to behaviorists, Gestalt psychologists proposed that organized structures, i.e., Gestalts, form the fundamental units of visual processing (rev., Wagemans, 2015). Hence to define a stimulus, one needs to refer to the internal cognitive structures of the subject. Gestaltists proposed that grouping operations can generate a variety of possible groups, starting from simple ones and proceeding towards more complex ones, until the structure meets the goals of the observer. They introduced the concept of *attitude* to express this active role of internal structures in giving a directness to perception, such as the expectation of a particular organization or outcome. Attention, in turn, was defined as a special case of attitude, which is *unspecific* towards a particular organization or outcome. These ideas found support in experiments that employed decoy tasks to minimize the involvement of attention (Krechevsky, 1938; Kohler & Adams, 1958).

The relationship between perceptual grouping and attention was investigated by Beck (1966a,b; 1967) who proposed that grouping occurs before the deployment of attention, a view supported by others (Caelli & Julesz, 1979; Julesz, 1991). On the other hand, Treisman and colleagues suggested that grouping and binding of stimuli occur pre-attentively *within* feature dimensions, but they require attention to occur *across* feature dimensions (Treisman, 1982; Treisman & Gelade, 1980). One pitfall of controlling attention by decoy/cover tasks is that it relies on memory for report. Hence, these studies may conflate the role of attention in memory with its role in perceptual grouping. Moore and Egeth (1997) used an implicit-measure approach to circumvent this problem and provided evidence that grouping occurs pre-attentively.

Although these results may appear contradictory, the view proposed by Gestalt psychologists, that attention and grouping are different but interacting processes, are consistent with these findings. Figure 1 shows some simple examples to illustrate this concept. In Figure 1a, the figure-ground organization is bistable in that both organizations (faces and vase) are of similar strength. Focusing attention can alter our percept. The hexagon in Figure 1b is contained in Figure 1c; however, no matter how much attention is directed into elements in Figure 1c that make up the hexagon, one cannot perceive the hexagon as a figure in itself in Figure 1c.

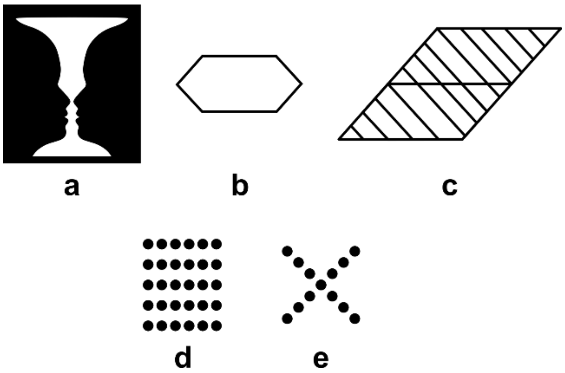


Figure 1. Examples of different grouping organizations. Reprinted with permission from Aydın, Herzog, & Öğmen (2011). Copyright (2011), Elsevier.

If attention and grouping are different but interacting processes, it remains to determine how they interact. According to Gestalt psychology, elements are grouped into wholes by a variety of laws such as common fate, similarity, proximity, good continuation, etc. But what holds elements of a group together? We suggest that each group possesses a reference-frame according to which elements are synthesized into groups. A simple example of this is the motion of a reflector on a bicycle wheel. In isolation, it is perceived to move on a cycloidal trajectory, which corresponds to its retinotopic motion when the eyes do not follow the reflector, but are directed at a stationary fixation mark. However, when a second reflector is added at the center of the wheel, the two reflectors become part of the same group and the trajectory of the reflector appears circular (e.g., Proffitt & Cutting, 1980; Proffitt, Cutting, & Stier, 1979). We perceive circular motion because the two reflectors on the wheel are grouped into the same Gestalt and the motion of the reflector at the center of the wheel serves as a *reference-frame* for the group. At the same time, it is impossible to perceive the cycloidal motion because this “non-retinotopic” reference-frame dominates (Boi, Öğmen, Krummenacher, Otto, & Herzog, 2009; Lauffs, Choung, Öğmen, & Herzog, 2018; Lauffs, Öğmen, & Herzog, 2017). Hence, reference frames determine how stimuli are grouped and what we perceive consciously.

To capture these ideas in a simple model, we proposed a “two-stage” spatio-temporal processing architecture (Figure 2), in which the first stage consists of grouping stimuli by Gestalt principles with modulation from attentional processes in order to establish reference frames for each group (Ogmen & Herzog, 2010, see Figure 18). These reference frames are then used to attribute features according to group membership and synthesize “objects” or “Gestalts” of perception.

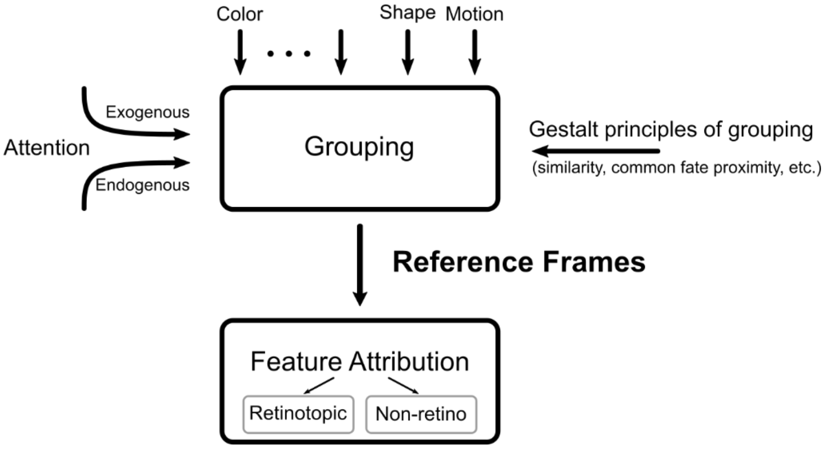


Figure 2. The two stage model: Features such as shape, motion, and color are used along with Gestalt principles of grouping, such as similarity, common fate and proximity, to form perceptual groups. A reference-frame is determined for each group and stimulus features are attributed to each group according to these reference frames. The retinotopic or the non-retinotopic percept results from differences in the attribution of features.

Attention can play a role in this process for example by modulating the perceptual groups, in particular when they are ambiguous, or by directly establishing a reference-frame that follows the spatio-temporal trajectory of focal attention.

The importance of grouping is evident when considering that vision is retinotopic in the first stages of vision. For example, neighboring elements in the visual field activate neighboring photoreceptors in the retina. This retinotopic encoding principle is preserved in the LGN and early visual areas. However, as shown by the bicycle example above, perception is usually non-retinotopic. Hence, the grouping operations at the first stage establish not only retinotopic reference-frames, but also non-retinotopic reference frames.

The Ternus-Pikler display is a suitable way to investigate the transition from retinotopic to non-retinotopic motion perception. When two disks are briefly presented on a computer screen and reappear at the same position after an inter-stimulus interval (ISI), observers perceive two flickering disks. In the Ternus-Pikler display, a third disk is added alternately to the left or right to induce apparent motion. When the ISI is very brief (e.g., 0 ms) the third disk appears to jump from the left to right of the two stationary disks, which is referred to as element motion. When the ISI is long (e.g., 200 ms), the three disks form a perceptual group and all three disks appear to shift left and right in concert, which is referred to as group motion. In past experiments (e.g., Boi et al., 2009), we used this effect to study non-retinotopic motion perception. We added a white dot to each disk, which we repositioned from frame to frame (Figure 3). The observer’s gaze was focused on a central fixation point and eye fixation was monitored by an eye tracker. Hence, the stimulus positions were the same in screen-based and retinotopic coordinates. We chose the dot positions so that the dots in the two stationary disks appeared to move up-down or left-right when no group or element motion was perceived. In the following, this linear dot motion is referred to as the retinotopic dot motion because it describes the dot's motion in retinotopic coordinates (Figure 3a and b). When the ISI was long enough for group motion, the dot in the middle disk was perceived to rotate. Because the dot did not rotate in retinotopic coordinates, dot rotation had to be computed non-retinotopically after group motion was established. The perceived dot rotation is therefore referred to as non-retinotopic dot motion (Figure 3c). In other words, the grouping of elements across the two frames of the Ternus-Pikler display determines element correspondence across the two frames (see the arrows in Figure 3) which in turn serves as reference-frames according to which the dot rotation is perceived. In retinotopic coordinates, the dots still move linearly up-down and left-right. Here, we used a modified version of the Ternus-Pikler display (Lauffs et al., 2018) where the dots rotated in retinotopic coordinates (Figure 3d), instead of linear movement. Therefore, both retinotopic and non-retinotopic dot motion included either clockwise or counter clockwise rotation.

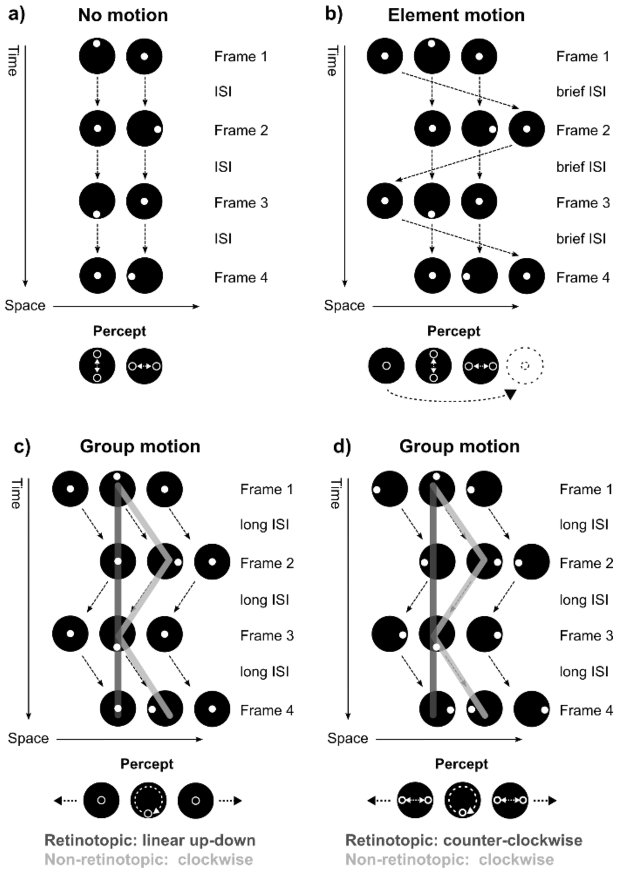


Figure 3. Ternus-Pikler display with linear retinotopic and circular non-retinotopic motion (Boi et al., 2009). a) No motion condition. When two disks are presented on the screen, the dot in one disk is perceived to move up-and-down in one and left-and-right in the other disk. b) When a third disk is added alternately to the left and right with a short ISI (e.g., 0 ms), the disks are perceived as two flanking disks in the middle and one disk jumping left-and-right ("element motion"). Dots in two middle disks are perceived to move as in the two disks condition, while the dot in the third disk stays in the center (retinotopic percept). c) When the ISI is prolonged (e.g., 200 ms), the three disks are perceived to move left-and-right in concert ("group motion"), and the dot motion percept changes: The dot in the middle disk is perceived to rotate (non-retinotopic percept), and the dots in the left and right disks are perceived to stay in the center. d) In the current experiment, circular retinotopic motion, instead of linear retinotopic motion, is used as in Lauffs et al. (2018). Therefore, both the retinotopic and non-retinotopic interpretation included either clockwise or counter clockwise rotation.

In previous studies (Boi et al., 2009; Lauffs et al., 2018; Lauffs et al., 2017), it was hypothesized that the grouping of disks across the two frames determines the reference-frame for evaluating the motion of the dots. It was also shown that attention can modulate spatio-temporal grouping in Ternus-Pikler displays, with group motion requiring more attentional resources to prevail against element motion (Aydin, Herzog, & Ogmen, 2011). Hence, an alternative account is possible. As shown in Figure 2, attention can play a role in determining the reference-frame either indirectly by modulating spatio-temporal groupings, especially when they are ambiguous, or directly by setting a reference-frame that follows its tracking trajectory. The latter idea is supported by studies showing a strong link between attentional tracking and the perception of apparent motion.

Two seminal studies showed that attentional tracking of salient texture elements (Lu & Sperling, 1995) or isoluminant color (Cavanagh, 1992) results in perceived motion of the tracked feature. Verstraten, Cavanagh, and Labianca (2000) measured the maximal speed of attentional tracking using bistable apparent motion stimuli. When confronted with bistable apparent motion, observers may decide to track a designated part of the stimulus, which disambiguates the apparent motion and results in a fixed perceived direction of motion. For instance, when a cross (x) and a plus sign (+) are presented in rapid alternation, observers perceive randomly reversing clockwise or counterclockwise rotation. However, when they track a spoke of the cross, a consistent direction of motion is perceived. In this case, the experimenter may evaluate the precision of the attentional tracking by asking the observer to compare the position of a probe to the position of the tracked stimulus after various intervals. It was observed that the maximal rate at which observers could track a rotating stimulus with an accuracy of 75% correct was 4-8 Hz. That is, tracking was not so much limited by the angular velocity of the spoke, but rather by the flicker rate of the apparent motion stimulus.

The flicker rate in studies on non-retinotopic motion perception in the Ternus-Pikler display is well within the limits of attentional tracking. For instance, group motion occurs in the Ternus-Pikler display with a stimulus-onset-asynchrony (SOA) of 333 ms, which corresponds to a flicker rate of 3 Hz (Lauffs et al., 2018). Thus, the rate of apparent motion is slow enough to allow for attentional tracking. It is therefore possible that attentional tracking of the central position in the Ternus-Pikler display during group motion contributed to the perception of non-retinotopic dot motion. The overarching goal of the present study was to investigate how multiple features, such as motion, shape and color, interact with attention to determine retinotopic or non-retinotopic reference frames. To evaluate the role of attentional tracking, we compared the perception of rotational dot motion in conditions where the relevant dot was highlighted by relative position (group motion), absolute position, color, or shape. All cues are expected to facilitate the tracking of the relevant motion stimulus, allowing us to evaluate the contribution of attentional tracking to the motion percept.

Experiment 1

In Experiment 1, we asked whether the perception of the retinotopic and non-retinotopic dot motion could be enhanced by showing the relevant disk in a distinct color. Through its salience, color may attract and direct attention whose spatio-temporal trajectory provides a reference-frame for motion perception (see left side of model in Figure 2). On the other hand, Hein and Moore (2012) demonstrated that color may also affect grouping by establishing spatio-temporal correspondence across stimulus frames (see right side of model in Figure 2). Thus, effects of color could result from attentional tracking or spatio-temporal correspondence. Experiment 1 deliberately accepts the ambiguity to establish effects of color, whereas Experiment 3 investigates conditions where the disks in the critical conditions were equally salient so that effects of attentional tracking could be isolated.

We expected beneficial effects of the color cue in conditions where performance was poor (see conditions C4 and C6 in Figure 4). Perception of retinotopic motion was poor when group motion prevailed with three disks, but excellent when only two stationary disks were shown (see conditions C4 and C2 in Figure 4). Possibly, group motion replaced the retinotopic reference-frame by a non-retinotopic one thereby making retinotopic motion invisible (Lauffs et al., 2018). For the perception of non-retinotopic motion, the situation was opposite: Performance was poor when the two disks were stationary and excellent when a third disk was added and group motion was perceived (see conditions C6 and C8 in Figure 4).

With group motion, a colored disk at a fixed retinal position (condition C3 vs. C4) may improve perception of retinotopic dot motion either because perceived group motion is reduced (Hein & Moore, 2012) or because attentional tracking of the relevant disk is facilitated. Conversely, a colored disk at the center position in the group (condition C7 vs. C8) may improve perception of non-retinotopic dot motion either because perceived group motion is increased (Hein & Moore, 2012; Proffitt & Cutting, 1980) or because attentional tracking of the central disk is facilitated. With two stationary disks, perception of non-retinotopic motion may improve because color establishes spatio-temporal correspondence of the relevant disk or because the attentional cue facilitates attentional tracking (condition C5 vs. C6). Importantly, color allows establishing spatio-temporal correspondence of the response-relevant disk by linking the most salient stimulus across frames. This process requires very little attention, but may be accomplished in a bottom-up manner (Itti & Koch, 2001).

Methods

Participants. Twelve participants took part in the experiment after giving written informed consent. Two participants were excluded from analysis, because their performance was lower than 60% correct in the retinotopic condition with two disks or the non-retinotopic condition with three disks (where excellent performance is expected based on previous experiments, Lauffs et al., 2018). We retained the data of 10 participants (mean age 22.6 ± 2.7 years, half were female, all were right-handed, and eight had right eye dominance). Seven participants had taken part in experiments with the Ternus-Pikler display in the past, but all were naïve to the purpose of the current experiment. All participants had normal or corrected to normal visual acuity, as indicated by a binocular score greater 1.0 in the Freiburg Visual Acuity Test (Bach, 1996). All experiments were conducted in accordance with the Declaration of Helsinki (World Medical Organization, 2013) and were approved by the local ethics committee.

Apparatus. Stimuli were displayed on a gamma-calibrated 24.5 inch BenQ XL 2540B LCD monitor (1920 x 1080 pixels, 60Hz, http://display-corner.epfl.ch). Viewing distance was 66 cm. The participants' chin and forehead were positioned in a SMI iViewX Hi-Speed 1250 eye tracker (Sensomotoric Instruments, Teltow, Germany), which was used to monitor eye fixation. Sample rate was 500 Hz and binocular samples were averaged to reduce noise. Trials containing eye movements larger than 1.5° (degrees of visual angle) or periods of data loss longer than 250 ms were discarded and repeated at a randomly chosen moment in the same block. Responses were collected using hand-held push buttons. When no response was registered within 3 secs or an eye movement was detected, the trial was repeated at a randomly chosen moment in the same block. A feedback tone sounded when no response was registered. In the case of eye movements, a feedback tone was played and a text message was displayed, reminding the participant to keep their gaze on the fixation point. No error feedback was provided.

Task, procedure, and design. The eight experimental conditions (C1-C8) are illustrated in Figure 4. We presented one block of 32 trials for each of the eight conditions. The order of blocks with color cue was random, but blocks with a colored disk were always followed by a block of the corresponding condition without color cue (i.e., C1 was followed by C2). Observers were instructed before each block of trials, using a video animation of the upcoming stimulus (provided in the Supplementary Materials). In a block with a color cue, participants were instructed to report the rotation of the dot (clockwise vs. counterclockwise) of the colored disk. In blocks without color cue, the position of the task-relevant disk was the same as in the preceding block *with* color cue. That is, observers were instructed to track the disk that appeared on the left-hand side in the first frame, and whether it would change position every other frame or not. Thus, there was no ambiguity regarding the response-relevant disk when all disks were black. In each trial, either two or three disks with white dots were presented (see Figure 4 and Movies 01 - 08 in the Supplementary Materials). In the conditions with the color cue, one disk was turquoise and the other disk(s) were black. Otherwise, all disks were black. The turquoise disk was either presented at the same position on the screen in all frames of the trial (retinotopic conditions) or switched position with a neighboring black disk in every other frame (non-retinotopic conditions). In each block of trials, the retinotopic and non-retinotopic rotation directions (clockwise vs. counterclockwise) and the initial orientation of the relevant rotation (3, 6, 9, 12 o'clock) were counterbalanced in a fully factorial fashion. The retinotopic and non-retinotopic rotation could hence be in the same (e.g., both clockwise) or in opposite directions (i.e., one clockwise and the other counter-clockwise).

Before the experiment, participants performed one training block of 16 trials with auditory error feedback, and one warm-up block of 48 trials without error feedback. In these blocks, we only used the stimuli where one of the disks was turquoise and presented the different conditions in random order. The participants were instructed to report the rotation of this disk after presentation of the last frame.

Stimulus. Either two or three disks with a diameter of 2° were presented 4° above a central fixation point (diameter = 0.05°, red, 20 cd/m^2). The disks were horizontally aligned and separated by a gap of 0.5°. Each disk contained a white dot (100 cd/m^2) with a diameter of 0.25°. A white dot was positioned either in the center of the disk or halfway between the center and the rim in different angular positions (3, 6, 9, or 12 o'clock). The disks were either black (0.4 cd/m^2) or turquoise (33 cd/m^2) depending on the condition. The background was midlevel grey (50 cd/m^2).

The disks were presented for 100 ms and reappeared after an inter-stimulus interval (ISI) of 200 ms. Per trial, only four stimulus frames with disks and dots were presented, preceded by two frames and followed by one frame, in which the disks were presented without dots. In trials with two disks, the disks were presented in the same positions in all frames. In trials with three disks, a third disk with a white dot was added alternately to the left and right. The disks were then perceived to move as a coherent group, alternately to the left and right, by exactly one inter-stimulus distance (2.5°). When only two disks were presented, the disks appeared to flicker at the same position.

From frame to frame, the dots were displaced within the disks to induce apparent motion. The perceived dot motion depended on the number of disks and the motion type (retinotopic, non-retinotopic) that was presented. When three disks were presented, the dot in the middle disk was perceived to rotate and the dots in the flanking disks were perceived to jump up-down or left-right in every second frame (*non-retinotopic* percept). When two disks were presented, the dot in the left disk was perceived to rotate and the dot in the right disk was perceived to jump up-down or left-right in every second frame (*retinotopic* percept). Importantly, the dot positions were identical in the two and three disk conditions (as introduced in Lauffs et al., 2018). The addition or omission of the third disk changed whether the dot motion was perceived in retinotopic or non-retinotopic coordinates, by changing the perceptual organization (i.e., whether the disks are perceived as moving left-right or stationary).

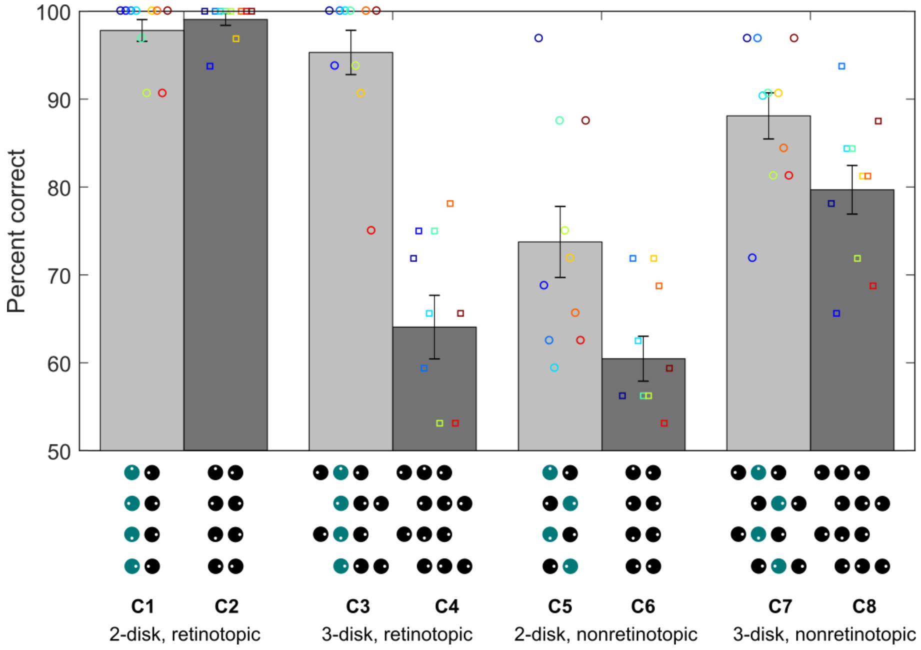


Figure 4. Dot rotation direction discrimination in Experiment 1. C1 and C2) Retinotopic tracking with 2 disks: Two disks are presented and participants were asked to track the retinotopic rotation. C3 and C4) Retinotopic tracking with 3 disks: Three disks are presented and participants were asked to track the retinotopic rotation. C5 and C6) Non-retinotopic tracking with 2 disks: Two disks were presented and participants were asked to track the non-retinotopic rotation. C6 and C8) Non-retinotopic tracking with three disks: Three disks were presented and participants were asked to track the non-retinotopic rotation. Error bars are ± 1 SEM. Colored data points show the mean performance of the individual observers.

Results

We calculated the mean percentage of correct responses for each of the eight conditions. Group and individual means are shown in Figure 4. We subjected individual means to a 2 (dot motion: retinotopic, non-retinotopic) x 2 (number of disks: 2, 3) x 2 (color cue: present, absent) analysis of variance (ANOVA) and found a significant three-way interaction, *F*(1, 9) = 61.23, *p* < .001, which justified separate two-way ANOVAs for conditions with retinotopic and non-retinotopic dot motion. Other effects were also significant in the three-way ANOVA, but are not reported for brevity.

A 2 (number of disks: 2, 3) x 2 (color cue: present, absent) ANOVA on percent correct for judgments of retinotopic dot motion found a main effect of the number of disks, *F*(1, 9) = 61.71, *p* < .001, and presence of a color cue, *F*(1, 9) = 79.45, *p* < .001. These main effects were modulated by a significant interaction, *F*(1, 9) = 89.8, *p* < .001. Inspection of conditions C1-C4 in Figure 4 suggests that judgments of retinotopic dot motion were poor when group motion was perceived (64% in C4), but highly precise in the remaining conditions (> 95% in C1, C2, and C3). Possibly, the adoption of a reference-frame based on group motion made the judgment of response-relevant retinotopic position difficult (condition C4), but when the color cue indicated the response-relevant retinotopic position (condition C3) this retinotopic cue was sufficient to alter the reference-frame into a retinotopic one. Paired *t*-tests confirmed that performance was better with than without color cue (condition C3 vs. C4, 95% vs. 64%), *t*(9) = 10.35, *p* < .001. The perception of retinotopic dot motion with color cue was indistinguishable from the perception of retinotopic dot motion without color cue (condition C1 vs. C2, 95% vs. 99%), *p* = .168.

Inspection of individual means in Figure 4 suggests that there was a ceiling effect in conditions C1, C2, and C3, which may compromise the normality of the data. By Kolmogorov-Smirnov test, we confirmed that the data were not normally distributed in these conditions, *p*s < .004. Therefore, we replaced the above *t-tests* with a nonparametric test (related-samples Wilcoxon signed rank test), but found the results to be unchanged.

The same 2 x 2 ANOVA was also performed on individual percent correct in the conditions with non-retinotopic dot motion (conditions C5-C8 in Figure 4). Performance was better with 3 than with 2 disks (84% vs. 67%), *F*(1, 9) = 72.99, *p* < .001, suggesting that group motion and its attendant non-retinotopic reference-frame helped perceive the non-retinotopic dot motion. Further, the main effect of color cue, *F*(1, 9) = 11.31, *p* = .008, showed that the color cue improved performance (80% vs. 70%). There was no interaction, *p* = .313, suggesting that the differences in the effect of color, which are apparent in Figure 4, were not reliable. Separate paired t-tests confirmed that the color cue improved performance in the 2-disk condition where performance was initially poor (60% vs. 74%), *t*(9) = 2.47, *p* = .036 , but also in the 3-disk condition where performance was initially good because of group motion (80% vs. 88%), *t*(9) = 5.67, *p* < .001.

Discussion

We measured the effect of a color cue on the perception of retinotopic and non-retinotopic dot motion. Whereas the retinotopic disk motion was masked by group motion in the same color condition, adding the color cue led to a strong increase in performance (see conditions C3 and C4 in Figure 4), suggesting that color either reduced the conflicting group motion (Hein & Moore, 2012) or facilitated the attentional tracking of retinotopic motion. For *non-retinotopic* dot motion, color improved performance in the condition benefitting from the intuitive perception of group motion (condition C7 vs. C8). The improvement may result either from improved perception of group motion (Hein & Moore, 2012) or from enhanced attentional tracking. Finally, color also improved perception of non-retinotopic motion with two stationary disks (condition C5 vs. C6). The condition with two stationary disks relies exclusively on attentional tracking because no other cue is available. With a colored disk, it is likely that attentional tracking of dot motion was facilitated. However, it may also be that the colored disk helped to establish spatio-temporal correspondence across stimulus frames.

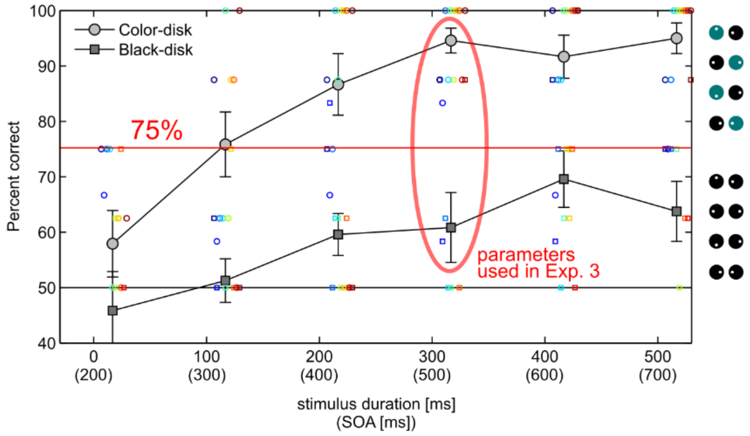
Experiment 2

In Experiment 1, we showed that perception of both the retinotopic and non-retinotopic rotation improved with color cues. In Experiment 2, we evaluated the timing parameters that lead to the maximal benefit of color before we ran the critical comparisons in Experiment 3. In one condition, we parametrically varied the stimulus duration and kept the ISI duration constant. In another condition, we varied the ISI duration and kept the stimulus duration constant.

Methods

The methods for Experiment 2 were identical to Experiment 1 with the following exceptions. We only used the 2-disk stimulus and instructed observers to track the non-retinotopic dot rotation, starting with the left disk in the first frame (see Figure 5 and Movies 08 - 32 in the Supplementary Materials). In the condition with fixed ISI, the stimulus duration was randomly varied from trial to trial (17 - 517ms in 100 ms steps) and the ISI was fixed at 200 ms. In the condition with fixed stimulus duration, the ISI duration was randomly varied from trial to trial (0 - 500ms in 100 ms steps) and the stimulus duration was fixed at 100 ms. For comparison with research by Verstraten et al. (2000), we also calculated the SOA (= stimulus duration + ISI). The SOA varied from 217 - 617 ms with fixed ISI and from 100 - 600 ms with fixed stimulus duration. The different stimulus/ISI durations and the directions of the retinotopic and non-retinotopic rotations (clockwise, counterclockwise) were used in a balanced 6x2x2 factorial design and presented in random order. The rotation started randomly in a 3, 6, 9, 12 o'clock orientation with equal probability. Per condition, 48 trials were presented (i.e., 8 trials per stimulus duration/ISI). Each condition was first run with a color cue, followed by an identical block without color cue. The order of conditions was counterbalanced across participants.

a) ISI = 200ms, stimulus duration variable



b) stimulus duration = 100ms, ISI variable

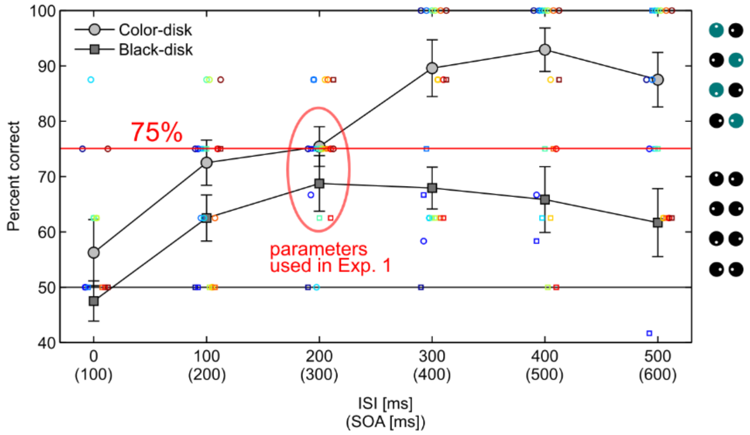


Figure 5. Performance with non-retinotopic dot rotation in Experiment 2. We varied either the stimulus duration (a) or the ISI (b). When the tracked disk was distinctly colored, performance improved from chance-level to near-perfect performance. When both disks were black, performance did not exceed 70% correct. Colored data points represent individual participants' performance. Error bars represent one standard error of the mean.

Results and Discussion

The ability to correctly indicate the non-retinotopic dot rotation increased with ISI and stimulus duration (Figure 5). Because the SOAs were not the same in conditions with fixed ISI and fixed stimulus duration, and SOA is key to attentional tracking of apparent motion (Verstraten et al., 2000), separate ANOVAs were carried out.

A 2 (tracking cue: color, none) x 6 (stimulus duration: 17, 117, 217, 317, 417, 517 ms) on individual means from the condition with fixed ISI of 200 ms showed that performance was better with tracking of the color cue compared to tracking without the color cue (84% vs. 58%), $F(1, 9) = 70.56$, $p < .001$. The effect of stimulus duration, $F(5, 45) = 13.86$, $p < .001$, showed that performance increased from 52% at the shortest to 79% at the longest duration. The interaction was not significant, $p = .33$, suggesting that the rate at which performance increased with increasing stimulus duration (the slope of the curves in Figure 5a) was not reliably different between the two tracking conditions.

Another 2 (cue: color, none) x 6 (ISI: 0, 100, 200, 300, 400, 500 ms) on individual means from the condition with fixed stimulus duration of 100 ms (see Figure 5b) showed better performance with one colored disk than with two black disks (79% vs. 62%), $F(1, 9) = 14.92$, $p = .004$, and increasing performance with ISI (from 52% at the shortest to 75% at the longest ISI), $F(5, 45) = 12.51$, $p < .001$. Additionally, there was an interaction of cue and ISI, *F*(5, 45) = 2.97, *p* = .021, confirming that the difference between color and no cue condition increased from 8% at the shortest ISI to 26% at the longest ISI.

These results indicate that a color cue improved the perception of non-retinotopic dot motion. Without color cue, however, performance did not exceed 70%, indicating that perception of non-retinotopic motion was severely limited, even with long periods of time between stimulus frames.

In Figure 5, we added a horizontal line to mark 75% correct responses, which were used as threshold in Verstraten et al. (2000). For the condition with color, the stimulus duration corresponding to 75% correct responses was around 100 ms (i.e., SOA of 300 ms, see Figure 5a). The ISI corresponding to 75% correct responses was around 200 ms (i.e., SOA of 300 ms, Figure 5b). Thus, we estimate the rate of apparent motion resulting in 75% correct to be roughly 3 Hz, which is at the lower end of the tracking limits reported in Verstraten et al. (2000).

Concerning the selection of an SOA that maximizes differences between attentional tracking with and without color cue, the experiment did not provide a clear answer. In particular, there was no statistical evidence for an interaction between time interval and color cue when the ISI was fixed (see Figure 5a). For lack of a better criterion, we selected the interval with the descriptively largest difference between conditions, the stimulus duration of 300 ms and the ISI of 200 ms (i.e., SOA of 500 ms). Finally, it is noteworthy that tracking of non-retinotopic motion without any external cue (i.e., with the black dots) was never better than 70%. Thus, the attentional selection of alternating horizontal stimulus positions was poor, but could be improved when color established spatio-temporal correspondence.

Experiment 3

In Experiment 3, we investigated the perception of non-retinotopic dot motion in more detail. Experiment 1 showed that making the response-relevant disk salient by means of a color cue improved the perception of non-retinotopic dot motion (see conditions C5-C8 in Figure 4). Because the salient color cue also changed spatio-temporal correspondence, improved performance could not be attributed to attentional tracking alone. In Experiment 3, the conditions of interest featured two colored disks of equal saliency (with physically isoluminant colors). Thus, attentional selection of one of the two colors was necessary. In contrast, it was no longer possible to establish spatio-temporal correspondence based on saliency. Rather, the response-relevant, but inconspicuous color had to be attentionally tracked. In another condition, we presented two equally salient shapes (with equal surface area) instead of two colors. The central question was whether attentional tracking of color or shape would improve perception of non-retinotopic motion relative to the condition without external cues. Further, we asked whether non-retinotopic motion perception would reach similar levels for salient stimuli, requiring little attentional tracking, as for inconspicuous elements that depended on attentional tracking. A single dot on two stationary disks and a single dot on a single disk were used to induce spatio-temporal correspondence by saliency. We refer to these cues as luminance-defined cues. Finally, we repeated the condition with group motion to evaluate whether attentional tracking or luminance-defined cues allow for similar levels of non-retinotopic motion perception.

Methods

The methods for Experiment 3 were identical to Experiment 1, unless noted otherwise. We used a stimulus duration of 300 ms and an ISI of 200 ms. Ten new, naïve observers participated in the experiment (mean age: 21.9 years, *SD* = 2, range: 19 - 25, 6 female, 9 right handed, 7 right eye dominance). All observers had normal color vision as tested with the Ishihara test for color deficiency (Ishihara, 2004). Observers were instructed to report the direction of the non-retinotopic dot rotation (clockwise vs. counterclockwise), starting with the left disk in the first stimulus presentation of each trial. We again used the movies provided in the Supplementary Materials to instruct the participants (see Figure 6 and Movies 33 - 45). The conditions were presented in random order, with one block of 32 trials each. Retinotopic and non-retinotopic rotation directions and the initial orientation of the rotating dot (3, 6, 9, 12 o’clock) were counter-balanced in a randomized full-factorial design for each observer and block. Before the experiment, each observer performed one practice block of 16 trials using the same stimulus as in Experiment 1. Auditory feedback for incorrect responses was provided only during the practice block.

Group Motion. Three black disks moved left and right in concert. Only the middle disk contained a white dot that performed a non-retinotopic rotation. This condition is the baseline condition and referred to as "3-disk" condition.

Luminance. In one condition, a single black disk moved left and right and contained a white dot that performed a non-retinotopic rotation ("1-disk"). In another condition, one of two disks had a rotating dot ("1-dot"). The dot was presented in the left disk in uneven numbered frames and in the right disk in even numbered frames.

Color. Participants tracked the color of the disk that was shown on the left position when the dot appeared. The color of the tracked dot remained the same in a block of trials. In one condition, each disk was enclosed by a square-shaped frame, which was blue for one and green for the other disk (physically isoluminant, both 45 cd/m^2). In another condition, the disks were not black, but one was blue and the other green. In a third condition, both disks were black, but one dot was blue and the other green.

Shape*.* Instead of two disks, we used one disk and one square. Both were black with white dots and had the same surface area. Participants tracked the shape on the left when the dots appeared. The shape stayed the same in a block of trials.

No cue. Two stationary black disks were presented. Both disks had white dots in all frames. In addition to the non-retinotopic rotation, a retinotopic rotation could be perceived in the left disk, and a retinotopic up-down or left-right dot motion in the right disk.

Mixed. This condition was similar to the luminance condition with a single disk, except that the left disk had an additional dot in the even numbered frames. These additional dots were positioned so that a retinotopic rotation of the left disk could be perceived.

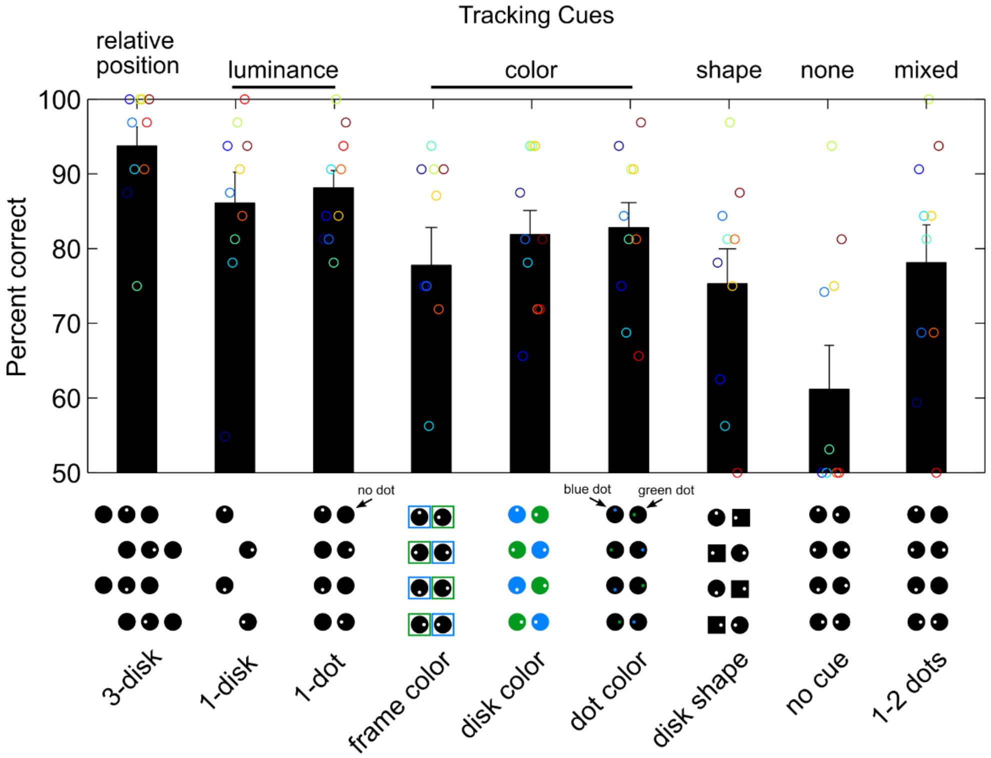


Figure 6. Discrimination of non-retinotopic dot rotation direction in Experiment 3. We varied the tracking cues to investigate whether attentional tracking of color or shape was as efficient as luminance-defined cues or group motion. Participants were asked to report the non-retinotopic rotation in all conditions. Error bars indicate one standard error of the mean. Colored circles depict the mean performance of the individual observers.

Results

The mean percentages of correct responses for the perception of non-retinotopic dot rotation are presented in Figure 6. We had created several versions of the luminance and color cue condition because we were unsure which would be most effective. To test for eventual differences, we performed a one-way ANOVA on the three color cue conditions, but found no significant effect, *F*(2, 18) = 1.48, *p* = .254. We therefore collapsed the color conditions. Similarly, we found no difference between the two luminance conditions, *t*(9) = 0.58, *p* = .576, and therefore collapsed these conditions as well. To evaluate effects of attentional tracking, we compared the color cue condition to the group motion and luminance cue conditions. Results from *t-tests* were confirmed by non-parametric Wilcoxon tests. Performance with color cues was worse than with group motion (81% vs. 94%), *t*(9) = 2.81, *p* = .021, but better than without cues (81% vs. 61%), *t*(9) = 3.71, *p* = .005. Importantly, there was no significant difference between color and luminance cues (81% vs. 87%), *t*(9) = 1.25*, p* = .244, showing that attentional tracking of color was as efficient as spatio-temporal correspondence established by luminance cues. Performance with attentional tracking of shape was non-significantly worse than tracking of color, *t*(9) = 2.13, *p* = .062. Possibly, it was more difficult to discriminate between the two shapes in peripheral vision than between the two colors. Further, we explored performance in a condition in which the presence and absence of external cues alternated (1-2 dots) and found performance (78%) to be in the range of performance with external luminance, color or shape cues (75-87%).

Discussion

Experiment 3 showed that the perception of non-retinotopic dot motion is best in the Ternus-Pikler display with group motion. Perception of non-retinotopic motion with other cues, such as luminance, color, or shape was worse, suggesting that group motion is a powerful cue to non-retinotopic motion perception.

We were interested in isolating attentional tracking in non-retinotopic motion perception. As in Experiments 1 and 2, performance in the no-cue condition was poor, showing that attentional tracking of position alone was difficult. However, the drop in performance from group motion to the no-cue condition may underestimate the efficiency of attentional tracking because external cues beyond stimulus position were missing. After all, group motion is established by the Gestalt principle of "common fate", which is a powerful cue beyond disk position, whereas only disk position was available in the no-cue condition. Therefore, we created conditions where observers had to attentionally track a color or a shape when two equally salient colors were available. While spatio-temporal correspondence could be established by the Gestalt principle of similarity, it was nonetheless necessary to select one of the two colors to accomplish the task. Thus, the color conditions isolate attentional tracking, but provide an external cue beyond disk position. We found that attentional tracking worked as well as luminance-defined cues, showing that attentional tracking may contribute to non-retinotopic motion perception. However, performance with color (or shape) cues was worse than with group motion, suggesting that non-retinotopic motion perception cannot be entirely accounted for by attentional tracking. Further, performance with luminance-defined cues was also worse than with group motion. In the 1-disk condition, only a single disk was shown so that establishing spatio-temporal correspondence was easy. However, performance was actually worse than with group motion, suggesting that the reference frame created by the outer disks facilitated perception compared to a single object.

General Discussion

Elements are grouped into wholes by a variety of Gestalt laws such as common fate, similarity, proximity, good continuation, etc.The two-stage model that we consider here (Figure 2) proposes that each group is endowed by a reference-frame that guides the attribution of stimulus features according to group identities. For example, consider the phenomenon of crowding, which refers to the failure to recognize visual stimuli when flanked by other stimuli (Andriessen & Bouma, 1976). The target stimulus itself is visible but features of the target and adjacent stimuli (“flankers”) appear mixed up (Pelli, Palomares, & Majaj, 2004). However, we showed that when the flankers do not belong to the same group as the target, the target itself becomes easily identifiable (Herzog & Manassi, 2015; Herzog, Sayim, Chicherov, & Manassi, 2015; Saarela & Herzog, 2008). One possible explanation for this effect is that segregating the target and flankers into two distinct groups creates two different reference-frames so that features of the flankers are not attributed to the target and vice-versa.

For dynamic stimuli, perceptual grouping occurs in space *and* time and the resulting spatio-temporal reference frames determine how features are attributed within perceptual groups. In sequential metacontrast, stimuli are grouped into different motion streams and the processing of features of individual elements depend on group membership of the elements (Otto, Ogmen & Herzog, 2009; Herzog, Otto, & Ogmen, 2012; Otto, Ogmen, & Herzog, 2006). Vernier offsets of elements within the same group (motion stream) are integrated whereas Vernier offsets of elements in different groups are not, regardless of spatio-temporal proximity (Otto et al., 2006). Similarly, as shown here and in previous work (Boi et al., 2009; Lauffs et al., 2018), in the Ternus-Pikler display, which pits retinotopic reference-frames against grouping-based non-retinotopic reference frames, feature processing depends on spatio-temporal grouping and the attendant reference-frame.

As shown in Figure 2, the choice of reference frames can also be influenced by attentional mechanisms. Attention can modulate and alter perceptual groups, especially if they are ambiguous. Previous research suggests that a reference-frame can also be established based on the spatio-temporal trajectory of focal attention (Cavanagh, 1992; Lu & Sperling, 1995; Verstraten et al., 2000). Here, we investigated whether cues such as color, shape, or luminance facilitates motion perception on the to-be-tracked object. In Experiment 1, we focused on two conditions that typically result in poor performance. First, it is difficult to perceive the retinotopic dot when there is group motion (see condition C4 in Experiment 1, e.g., Boi et al., 2009; Clarke, Öğmen, & Herzog, 2016; Lauffs et al., 2018; Lauffs et al., 2017). However, despite being invisible, retinotopic motion may strongly interfere with perception of non-retinotopic motion (Lauffs et al., 2018). Second, it is hard to perceive non-retinotopic dot motion with two stationary disks (see condition C6 in Experiment 1). We found that color cues improved performance in both cases, suggesting that attentional tracking may contribute to the perception of retinotopic and non-retinotopic motion perception. However, an alternative account in terms of changes in the perception of group or element motion (Hein & Moore, 2012) cannot be ruled out.

One argument against a strong role of attentional tracking is that perception of non-retinotopic motion with two stationary disks, which relies exclusively on attentional tracking, is poor (condition C6 in Experiment 1, no-cue conditions in Experiments 2 and 3). It may be that the perception of retinotopic motion inside the two stationary disks interfered with tracking the non-retinotopic motion between the left and right position. When a color cue was added (condition C5 vs. C6 in Experiment 1, color cue conditions in Experiments 2 and 3), performance improved moderately, but was far from ceiling. Similarly, performance improved moderately when we added a color cue with non-retinotopic motion perception (condition C7 vs. C8 in Experiment 1). The latter improvement may arise from attentional tracking or an enhancement of the group motion. Whatever the exact mechanism(s) at work, the effects of the color cue are limited. As a further case in point, Experiment 3 showed that non-retinotopic motion perception was worse with any of the tracking cues compared to group motion. Thus, our experiments point to a limited role for attentional tracking, but cannot decide whether non-retinotopic motion perception with group motion arises from a motion processor after group motion is established or results directly from an attentional tracking. It was just recently found that even smooth motion percepts can result from tracking (Allard & Arleo, 2016).

The flicker rate in our paradigm was well within the limits of attentional tracking (Verstraten et al., 2000). Our results are surprising when put into the context of the existent literature on multiple object tracking (Cavanagh & Alvarez, 2005; Meyerhoff, Papenmeier, Jahn, & Huff, 2013; Pylyshyn & Storm, 1988; Vater, Kredel, & Hossner, 2016), where it has been shown that observers can track up to four disks over several seconds even when the disks cross the trajectories of a large number of distractor disks. In the light of powerful attentional tracking of objects undergoing smooth motion, it is surprising that observers can hardly track one disk in the 2-disk condition without the color cue (e.g., condition C6 in Experiment 1, no-cue condition in Experiments 2 and 3). This suggests that the spatio-temporal trajectory of focal attention in isolation plays a relatively weak role in the first stage of the two-stage model (Figure 2). Instead, interactions between attention and other grouping cues, seem to provide a much stronger basis for determining the reference-frame underlying stimulus processing.

Surprisingly, Experiment 2 showed that increasing the ISI or duration in this condition did not change the results substantially. Even when the ISI was 500 ms, attentional tracking was almost impossible when external cues were absent. In this case, the entire sequence lasted for 2.2 s, retinotopic motion still prevailed. With the addition of the color cue, ceiling performance was reached for SOAs of 400 to 500 ms. We suggest that built-in motion routines have a privileged role in establishing *spatio-temporal* groups. This leaves a limited role for attentional tracking for substantial times of processing. The primacy of motion in transforming retinotopic coordinates into non-retinotopic ones makes ecological sense. Due to our own movements and those of external objects, we receive highly dynamic stimuli according to retinotopic coordinates. Hence, motion is a relevant, abundant, and readily available feature. However, color and spatial cues also play an important role in perceptual grouping (Figure 2). They can override motion cues and allow tracking, especially when they are congruent with the task-dependent trajectory of attentional tracking. Whereas the role of attentional tracking is limited, we showed that attention operates after the reference frames are established, i.e. in non-retinotopic coordinates (Experiment 3; Boi et al., 2009; Boi, Vergeer, Öğmen & Herzog, 2011; Scharnowski, Hermens, Kammer, Öğmen, & Herzog, 2007).

Taken together, our results add evidence that attention and grouping are different but interacting processes. Our study highlights the importance of reference-frames and supports the two-stage model (Figure 2). The model clarifies how various features and cues can work in conjunction or in competition to determine prevailing groups. These groups in turn establish reference-frames according to which features are processed and bound together. Attention allows the selection from a variety of possible groupings, the grouping that fits best on the observer's internal state and goals, especially when Gestalt laws produce groups of similar strengths.

Acknowledgements

M.M. Lauffs, O.H. Choung, and M.H. Herzog were supported by the Swiss National Foundation 320030_176153 "Basics of visual processing: from elements to figures". D. Kerzel was supported by the Swiss National Foundation 100019_182146.

References

Allard, R., & Arleo, A. (2016). Position-based vs ener gy-based motion processing. *Journal of Vision*, *16*(12), 670-670.

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision research*, *16*(1), 71-78. doi:10.1016/0042-6989(76)90078-X

Allard, R., & Arleo, A. (2016). Position-based vs energy-based motion processing. *Journal of Vision, 16*(12), 670-670.

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: adverse interactions between line segments. *Vision Research, 16*(1), 71-78. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1258390

Aydin, M., Herzog, M. H., & Ogmen, H. (2011). Attention modulates spatio-temporal grouping. *Vision Research, 51*(4), 435-446. doi:10.1016/j.visres.2010.12.013

Beck, J. (1966a). Slant and shape variables in perceptual grouping. *Science*, *154*, 538-540.

Beck, J. (1966b). Effect of orientation and of shape similarity on perceptual grouping. *Perception & Psychophysics*, *1*(5), 300-302.

Beck, J. (1967). Perceptual grouping produced by line figures. *Perception & Psychophysics*, *2*(11), 491-495.

Boi, M., Öğmen, H., Krummenacher, J., Otto, T. U., & Herzog, M. H. (2009). A (fascinating) litmus test for human retino- vs. non-retinotopic processing. *J Vis, 9*(13), 5 1-11. doi:10.1167/9.13.5

Boi, M., Vergeer, M., Ogmen, H., & Herzog, M. H. (2011). Nonretinotopic exogenous attention. *Current Biology*, *21*(20), 1732-1737. doi:10.1016/j.cub.2011.08.059.

Caelli, T., & Julesz, B. (1979). Psychophysical evidence for global feature processing in visual texture discrimination. *JOSA*, *69*(5), 675-678. doi:10.1364/JOSA.69.000675

Cavanagh, P. (1992). Attention-based motion perception. *Science, 257*(5076), 1563-1565.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends Cogn Sci, 9*(7), 349-354. doi:10.1016/j.tics.2005.05.009

Clarke, A. M., Öğmen, H., & Herzog, M. H. (2016). A computational model for reference-frame synthesis with applications to motion perception. *Vision Research, 126*, 242-253. doi:10.1016/j.visres.2015.08.018

Hein, E., & Moore, C. M. (2012). Spatio-temporal priority revisited: The role of feature identity and similarity for object correspondence in apparent motion. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 975-988. doi:10.1037/a0028197

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. doi:10.1016/j.cobeha.2014.10.006

Herzog, M. H., Otto, T. U., & Ogmen, H. (2012). The fate of visible features of invisible elements. *Frontiers in Psychology, 3*, 119. doi:10.3389/fpsyg.2012.00119

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

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews: Neuroscience, 2*(3), 194-203. doi:10.1038/35058500

Julesz, B. (1991). Early vision and focal attention. *Reviews of modern physics*, *63*(3), 735. doi:10.1103/RevModPhys.63.735

Koffka, K. (1922). Perception: an introduction to the Gestalt-Theorie. *Psychological Bulletin*, *19*(10), 535.

Köhler, W., & Adams, P. A. (1958). Perception and attention. *The American journal of psychology*, *71*(3), 489-503.

Krechevsky, I. (1938). An experimental investigation of the principle of proximity in the visual perception of the rat. *Journal of experimental psychology*, *22*(6), 497. doi:10.1037/h0058982

Lauffs, M. M., Choung, O.-H., Öğmen, H., & Herzog, M. H. (2018). Unconscious retinotopic motion processing affects non-retinotopic motion perception. *Consciousness and Cognition, 62*, 135-147. doi:10.1016/j.concog.2018.03.007

Lauffs, M. M., Öğmen, H., & Herzog, M. H. (2017). Unpredictability does not hamper nonretinotopic motion perception. *Journal of Vision, 17*(9). doi:10.1167/17.9.6

Lu, Z.-L., & Sperling, G. (1995). Attention-generated apparent motion. *Nature, 377*, 237. doi:10.1038/377237a0

Meyerhoff, H. S., Papenmeier, F., Jahn, G., & Huff, M. (2013). A single unexpected change in target- but not distractor motion impairs multiple object tracking. *Iperception, 4*(1), 81-83. doi:10.1068/i0567sas

Moore, C. M., & Egeth, H. (1997). Perception without attention: Evidence of grouping under conditions of inattention. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 339. doi:10.1037/0096-1523.23.2.339

Öğmen, H., & Herzog, M. H. (2010). The geometry of visual perception: Retinotopic and nonretinotopic representations in the human visual system. *Proceedings of the IEEE*, *98*(3), 479-492. doi:10.1109/JPROC.2009.2039028

Otto, T. U., Ogmen, H., & Herzog, M. H. (2006). The flight path of the phoenix--the visible trace of invisible elements in human vision. *J Vis, 6*(10), 1079-1086. doi:10.1167/6.10.7

Otto, T. U., Ögmen, H., & Herzog, M. H. (2009). Feature integration across space, time, and orientation. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1670. doi:10.1037/a0015798.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12), 1136-1169. Retrieved from http://journalofvision.org/4/12/12/

Proffitt, D. R., & Cutting, J. E. (1980). An invariant for wheel-generated motions and the logic of its determination. *Perception, 9*, 435-449.

Proffitt, D. R., Cutting, J. E., & Stier, D. M. (1979). Perception of wheel-generated motions. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 289-302.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179-197.

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. doi:10.1167/8.3.23

Scharnowski, Frank, Frouke Hermens, Thomas Kammer, Haluk Öğmen, & Michael H. Herzog. "Feature fusion reveals slow and fast visual memories." *Journal of Cognitive Neuroscience* 19, no. 4 (2007): 632-641. doi:10.1162/jocn.2007.19.4.632

Tolman, E. C. (1932). The determiners of behavior at a choice point. *Psychological Review*, *45*, 1-41. doi:10.1037/0096-1523.8.2.194

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of experimental psychology: human perception and performance*, *8*(2), 194. doi:10.1037/0096-1523.8.2.194

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136. doi:10.1016/0010-0285(80)90005-5

Vater, C., Kredel, R., & Hossner, E.-J. (2016). Detecting single-target changes in multiple object tracking: The case of peripheral vision. *Attention, Perception, & Psychophysics*, 1-16. doi:10.3758/s13414-016-1078-7

Verstraten, F. A. J., Cavanagh, P., & Labianca, A. T. (2000). Limits of attentive tracking reveal temporal properties of attention. *Vision Research, 40*(26), 3651-3664. doi:10.1016/S0042-6989(00)00213-3

Wagemans, J. (2015). *The Oxford Handbook of Perceptual Organization*: Oxford University Press. doi:10.1093/oxfordhb/9780199686858.001.0001

2.2 Unconscious retinotopic information affects non-retinotopic motion perception

Full citation: Lauffs, M. M., **Choung, O. H.**, Öğmen, H., & Herzog, M. H. (2018). Unconscious retinotopic motion processing affects non-retinotopic motion perception. *Consciousness and cognition*, *62*, 135-147.

Summary:

In the TPD, dot motion is perceived relative to the disk motion. We only perceive the non-retinotopic dot motion. Here, we asked whether the invisible retinotopic dot motion is completely lost or survives, affecting the non-retinotopic perception.

We modified the TPD configurations, so that a dot rotation was perceived in both the retinotopic and non-retinotopic interpretations of the display. The retinotopic and non-retinotopic dot rotations could either be in the same direction (*congruent* rotation, e.g., both clockwise) or opposite directions (*incongruent* rotation; e.g., retinotopic clockwise and non-retinotopic counter-clockwise). We tested whether the retinotopic rotations, which were invisible when the group motion percept dominated, interacted with the non-retinotopic ones. We hypothesized that the incongruent retinotopic rotation might impair the non-retinotopic rotation percept, whereas congruent retinotopic rotation might facilitate it. We added one more invisible retinotopic rotation and asked whether one congruent and one incongruent retinotopic rotation cancel each other, and whether the detrimental effects of two incongruent retinotopic rotations would add up to even worse performance. Finally, we investigated whether visible *non*-retinotopic rotations affect the percept in a similar fashion as invisible retinotopic rotations.

We found three important observations. First, when one invisible retinotopic rotation was presented, only incongruent retinotopic rotation (which was in the opposite direction compared to non-retinotopic rotation) deteriorated the performance. However, congruent invisible retinotopic rotation did not improve the performance. Performance was equally good as in only non-retinotopic rotation condition (baseline condition). Second, when we added a second invisible retinotopic rotation to the stimulus, the performance pattern was similar to the one invisible retinotopic condition. When both retinotopic rotations were in the same direction as the non-retinotopic rotation, performance was equally good as in the baseline condition. When both retinotopic rotations were in the opposite direction of the non-retinotopic rotation, performance was even lower. Interestingly, when one of the two retinotopic rotations was in the opposite direction and the other in the same direction of the non-retinotopic rotation, two invisible retinotopic rotations did not cancel each other, instead only opposite retinotopic rotation deteriorated the performance. Therefore, congruent retinotopic rotation was ignored. These results show that there are strong effects of the invisible retinotopic motion on the visible non-retinotopic motion. However, these interactions are not linear, i.e., only incongruent rotation impacted. Finally, we presented non-retinotopic rotations in all three disks, but no retinotopic rotation. All three rotations were perceived, and observers reported the non-retinotopic rotation in the middle disk. Overall, performance differences were minor. These results strongly suggest that invisible (unconscious) rotation can impact the conscious non-retinotopic perception, but only when unconscious and conscious percepts are incongruent.

Postprint of the article published in *Consciousness and cognition*:

Unconscious retinotopic motion processing affects non-retinotopic motion perception

Marc M. Lauffs[a*], Oh-Hyeon Choung[a], Haluk Öğmen[b], & Michael H. Herzog[a]

[a] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, http://lpsy.epfl.ch, marc.lauffs@epfl.ch, oh-hyeon.choung@epfl.ch, michael.herzog@epfl.ch

[b] Department of Electrical and Computer Engineering, University of Denver, CO, USA, haluk.ogmen@du.edu

*Corresponding author

Highlights

- Visual motion is usually perceived in non-retinotopic, object-centered coordinates.
- The *retinotopic* motion trajectory is invisible, but it is processed unconsciously.
- Invisible retinotopic motion affects conscious non-retinotopic motion perception.
- The Ternus-Pikler display is a versatile new tool to study unconscious processing.

Abstract

Unconscious visual stimuli can affect conscious perception: An invisible prime can affect responses to a subsequent target. The invisible interpretation of an ambiguous figure can have similar effects. Invisibility in these situations is typically explained by stimulus-suppression in early retinotopic brain areas. We have previously argued that invisibility is closely linked to Gestalt ("object") organization principles. For example, motion is typically perceived in non-retinotopic, object-centered, and not in retinotopic coordinates. Such is the case for a bicycle-reflector that is perceived as circling, although its retinotopic trajectory is cycloidal. Here, we used a modified Ternus-Pikler display in which, just as in everyday vision, the retinotopic motion is invisible and the non-retinotopic motion is perceived. Nevertheless, the invisible retinotopic motion, can strongly degrade the conscious non-retinotopic motion percept. This effect cannot be explained by inhibition at a retinotopic processing stage.

Keywords

non-retinotopic processing; invisibility; ambiguous figures; consciousness

Introduction

Conscious and unconscious perception are usually investigated with experimental techniques such as backward masking (Bachmann, Breitmeyer, & Öğmen, 2007; Bachmann & Francis, 2013; Breitmeyer & Öğmen, 2006), binocular rivalry (Wheatstone, 1838; Blake, 2001), or (continuous) flash suppression (Wolfe, 1984; Tsuchiya & Koch, 2004). Interestingly, even when an element is invisible, it can influence the processing of visible elements. For example, an invisible prime can speed up the responses to an element presented later (Fehrer & Raab, 1962; Klotz & Wolff, 1995; Klotz & Neumann, 1999; Vorberg et al., 2003). Invisibility of the stimulus is usually explained by inhibition between neurons sensitive to the target and neurons sensitive to the mask or the stimuli in the other eye. In line with the organization of the early visual areas in cortex, inhibition is either explicitly or implicitly assumed to be *retinotopically* organized (Wandell, Dumoulin, & Brewer, 2008; Engel, Glover, & Wandell, 1997). For example, in binocular rivalry, neurons processing information originating from the left and right eye mutually suppress each other, by each inhibiting the neuron of the other eye coding for the same position in the visual field (Blake, 1989; however, see Leopold & Logothetis, 1996, Kovacs, Papathomas, Yang, & Fehér, 1996).

In ambiguous figures, one of two rivaling interpretations is perceived. For example, in Rubin's vase, the silhouette of either a goblet or two faces is perceived, depending on which one is interpreted as figure and which one as ground. Unlike in masking and rivalry, all elements of the display are fully visible, but one *interpretation* is suppressed. As in masking, the suppressed interpretation of the image can influence stimuli presented later (e.g., Peterson & Kim, 2001; Peterson & Skow, 2008). Also for ambiguous figures, retinotopic inhibition may explain why stimuli or interpretations are invisible, for example, by mutual inhibition of boundary-ownership neurons in V2 (Zhou, Friedman, & von der Heydt, 2000; Zhaoping, 2005).

Perception is usually *non-retinotopic*. The parts of a moving object are perceived relative to the object, rather than in retinal coordinates (Öğmen & Herzog, 2010; Duncker, 1929; Johansson, 1950, 1974, 1976; Clarke, Öğmen, & Herzog, 2016; Ağaoğlu, Clarke, Herzog, & Öğmen, 2016). For example, the reflector on the wheel of a bicycle is perceived to rotate, although its motion is cycloidal in retinotopic coordinates. The bicycle motion is perceptually discounted from the cycloidal retinotopic motion of the reflector. Only the circular non-retinotopic reflector motion is perceived consciously, whereas the cycloidal retinotopic motion is invisible. Similar to ambiguous figures, all elements are visible, only one interpretation is suppressed (Herzog, Hermens, & Öğmen, 2014). Obviously, the percept cannot be explained exclusively by retinotopic inhibition because it depends on non-retinotopic information. Almost nothing is known about the unconscious processing of invisible retinotopic motion and its influences on the consciously perceived non-retinotopic motion. Here, we used an adapted version of the Ternus-Pikler display (Ternus, 1926; Pikler, 1917). In the Ternus-Pikler display, two disks are briefly presented on a computer screen and reappear after an inter-stimulus interval (ISI). The disks are perceived to flicker on and off in the same positions (*no motion*). Next, a third disk is added alternately to the left or right. When the ISI is very brief (e.g., 0 ms) the third disk appears to jump from the left to right of the two stationary disks and so on (*element motion*). However, when the ISI is long (e.g., 200 ms), the three disks form a perceptual group and all three disks appear to shift left and right in concert (*group motion*).

In past experiments (e.g., Boi, Öğmen, Krummenacher, Otto, & Herzog, 2009), we used this effect to study non-retinotopic motion perception. We added a white dot to each disk, that we repositioned from frame to frame (Figure 1a). The observer's gaze was focused on a central fixation point and an eye tracker controlled that no eye movements were made. Hence, the stimulus positions were the same in screen based and retinotopic coordinates. We chose the dot positions, so that when no motion or element motion was perceived, the dots in the two stationary disks appeared to move up-down and left-right, respectively (*retinotopic dot motion percept;* Figure 1a left and Movie 1). However, when group motion was perceived, the dot in the middle disk was perceived to rotate (*non-retinotopic dot motion percept;* Figure 1a right and Movie 2). This dot rotation percept is *non-retinotopic*, because it can only be perceived in disk-centered coordinates; in retinotopic coordinates, the dots still move linearly up-down and left-right.

Interestingly, the perceptions of retinotopic and non-retinotopic dot motion are mutually exclusive: When two stationary disks are presented, the retinotopic dot motion is perceived, and the non-retinotopic dot motion is invisible. When the three disks in group motion are presented, the non-retinotopic dot motion is perceived, and the retinotopic dot motion is invisible. In both cases, the stimulus is identical on the screen and in retinotopic coordinates, except for the third disk. In both cases, the disks and dots are fully visible. Hence, the retinotopic dot motion does not become invisible because retinotopic information is suppressed, but because the addition of the third disk changes how the stimulus is grouped and interpreted. Here, we show that unconscious processing of the retinotopic motion strongly affects the conscious perception of non-retinotopic motion.

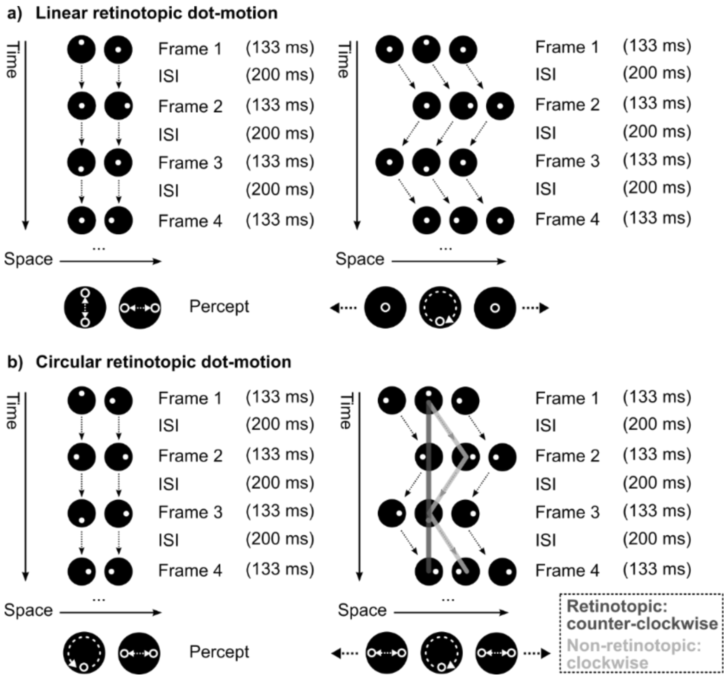

a) Linear retinotopic dot-motion
Time
Frame 1 (133 ms)
ISI (200 ms)
Frame 2 (133 ms)
ISI (200 ms)
Frame 3 (133 ms)
ISI (200 ms)
Frame 4 (133 ms)
Space
...
Percept
b) Circular retinotopic dot-motion
Retinotopic: counter-clockwise
Non-retinotopic: clockwise

Figure 1 a) Ternus-Pikler display with linear retinotopic and circular non-retinotopic dot motion (as introduced by Boi et al., 2009). Left: When two disks are presented, the dot in one disk is perceived to move up-and-down in one and left-and-right in the other disk (retinotopic percept). Right: When a third disk is added alternately to the left and right, the three disks are perceived to move left-and-right in concert (“group motion” percept for the disks) and the dot in the middle disk is perceived to rotate (non-retinotopic percept). In the left and right disk the dots are always in the center. b) Ternus-Pikler display with circular retinotopic and circular non-retinotopic dot motion. Left: When only two disks are presented, in the left disk the dot is perceived to rotate and in the right disk to jump left-and-right every second frame (retinotopic percept). Right: When a third disk is added alternately to the left and right, the dot in the middle disk is perceived to rotate and the dots in both outer disks jump left-and-right every second frame (non-retinotopic percept). Although the retinal image is identical apart from the third contextual disk, only the non-retinotopic dot motions is perceived with three disks. With two disks, only the retinotopic dot motion is perceived. We show here an example with clockwise non-retinotopic and counter-clockwise retinotopic rotation, but the rotations could be in both the same and opposite directions. Dark and light grey lines point out the retinotopic and non-retinotopic rotation, respectively, and were not presented during the experiment. Dots that are not involved in either rotation (first and third in Frame 1, third in frame 2, etc.) can be placed arbitrarily. We chose the depicted dot positions because the symmetry of the outer disks enhances the group-motion percept. Broken arrows indicate the perceived object correspondence and motion direction of the disks and were not presented during the experiment.

Experiment 1a: Interactions between retinotopic and non-retinotopic motions

In Experiment 1a, we chose the positions of the white dots so that a dot rotation was perceived in both the retinotopic and non-retinotopic interpretations of the display (cf. Figure 1b). When two disks were presented, retinotopic dot rotation was perceived in the left disk, while the dot in the right disk jumped left-and-right or up-and-down every second frame (Figure 1b left and Movie 3). When the third disk was added, non-retinotopic dot rotation was perceived in the middle disk, and the dots in the outer disks jumped left-and-right or up-and-down every second frame (Figure 1b right and Movie 4). The dot positions were identical in both cases, except that a third disk and dot were added. The retinotopic and non-retinotopic dot rotations could either be in the same direction (*congruent* rotation, e.g., both clockwise; Movie 5) or in opposite directions (*incongruent* rotation; e.g., retinotopic clockwise and non-retinotopic counter-clockwise; Movie 6). The observers were asked to report either the retinotopic or non-retinotopic rotation direction (clockwise/counter-clockwise).

To test the influence of the retinotopic rotation on the non-retinotopic percept we presented three disks and had observers report the non-retinotopic rotation direction. We then compared performance between trials in which the retinotopic rotation was in the same (Movie 5) versus the opposite direction of the non-retinotopic rotation (Movie 6). We hypothesized that incongruent retinotopic rotation might impair the non-retinotopic rotation percept, whereas congruent retinotopic rotation might facilitate it. As a baseline condition, we first presented a Ternus-Pikler display with only a non-retinotopic dot-rotation in the middle disk, but no dots in the flanking disks (Movie 7). To determine whether the influences on non-retinotopic motion are specific to the motion type, i.e., to determine whether the non-retinotopic rotation percept is modulated by retinotopic *rotation* rather than the mere presence of retinotopic dot-motions per se, we included control conditions in which the retinotopic dot motions were linear (up-down or left-right; Figure 1a and Movies 1-2) or random (Movie 8).

Methods

Observers. Sixteen naïve observers took part in the experiment, but three observers were excluded from the analysis due to inferior performance in the baseline condition (60% correct responses or less). Three other observers were excluded because they were unable to maintain stable central fixation. Hence, 10 observers were available for analysis (mean age 23.3 years, *SD* = 2.3 years, 5 female, 1 left-handed, 1 wore glasses). None of the observers had participated in earlier experiments with the Ternus-Pikler display and all were naïve to the purpose of the experiment. All observers had normal or corrected-to-normal visual acuity, as indicated by a binocular value ≥ 1.0 (corresponding to 20/20) in a program similar to the Freiburg visual acuity test (Bach, 1996). The experiments were approved by the local ethics committee and performed in accordance with the Declaration of Helsinki (World Medical Association, 2013). All observers gave written informed consent prior to the experiment. The observers were recruited from the EPFL student population and paid 20 CHF/h for their participation.

Apparatus. Stimuli were programmed in Matlab with Psychtoolbox (Brainard, 1997; Pelli, 1997) and presented on a 24.5 inch BenQ XL2540 LCD monitor (1920x1080 pixels, 60 Hz, http://display-corner.epfl.ch). Observers were positioned in the head rest of a SMI iViewX Hi-Speed 1250 eye tracker (Sensomotoric Instruments, Teltow, Germany) and viewed the stimuli from a distance of 0.66 m. Eye tracking data were recorded binocularly at 500 Hz and immediately averaged over both eyes to reduce noise. The room was well lit to facilitate the eye tracking.

Stimulus. Depending on the condition, either two or three black disks (0.3 cd/m^2) with a white dot (98 cd/m^2) were presented in each frame for 133 ms on a grey background (58 cd/m^2). Each stimulus frame was followed by an ISI of 200 ms. The black disks were 2.0° in diameter and separated horizontally by 0.5° (2.5° center-to-center). When three disks were presented, the stimulus shifted back and forth by one inter-stimulus distance (2.5°) per frame, so that two stimulus positions overlapped in all frames (cf. Figure 1). The white dots were 0.25° in diameter and positioned in the center of the disk or halfway between the disk's center and rim at the 3, 6, 9, or 12 o'clock position. The stimulus was presented two disk diameters (4.0°) above a central fixation point (red, $r = 0.025°$, 20 cd/m^2), whose horizontal position was midway between the two central disks. In each trial only four frames were presented, preceded by two and followed by one frame where the black disks were shown without the white dots. Trials were separated by an inter-trial interval of at least 1 sec, but started only after the observer fixated the fixation point.

In the 3-disks conditions with linear retinotopic motion, we randomized and counterbalanced whether the third disk first appeared to the left or right, the rotation direction of the dot in the middle disk (clockwise/counter-clockwise), and the dot position in the first frame of the trial (3, 6, 9, 12 o'clock). We presented all possible combinations of the above factors equally often and in random order. The stimulus was identical in the conditions with only two disks, except that the third disk was omitted. In each trial, the dot in the left disk moved either left-right or up-down with equal probability. In the first frame of the trial, the dot in the left disk randomly started in the disk center (50% of trials) or at a 3, 6, 9, or 12 o'clock orientation (12.5% of trials each). In the 3-disks conditions with circular retinotopic motion, we combined the retinotopic and non-retinotopic rotation directions (clockwise/counter-clockwise) in a factorial design and presented the different combinations in random order. The retinotopic and non-retinotopic rotations could hence be in the same (Movie 5) and in opposite directions (Movie 6) with equal probability. We randomized and counterbalanced whether the third disk first appeared to the left or right. The initial dot orientation for the retinotopic and non-retinotopic rotations (3, 6, 9, 12 o'clock) were each chosen randomly with equal probability. Again, the stimulus was identical in the conditions with two disks, except that the third disk was omitted. Detailed schematics of all possible stimulus configurations are provided in the online supplementary materials.

Depending on the condition, the observers were asked to report the non-retinotopic rotation direction (clockwise vs. counter-clockwise) or the retinotopic motion direction (clockwise vs. counter-clockwise, or up-down vs. left-right) via handheld push-buttons. The observers were re-instructed before each block with a demonstration movie of the stimulus they were about to see (the movies are provided in the supplementary materials). In the condition where three disks were presented and the (invisible) retinotopic rotation had to be reported, we explained the task by occasionally occluding the third disk to make the retinotopic rotation temporarily visible. In the condition where the two disks were presented and the (invisible) non-retinotopic rotation had to be reported, the experimenter pointed out the requested rotation with the mouse cursor, allowing the observers to track it. We ensured that the task was well understood, by asking for verbal reports of the requested rotation direction before we started the data collection. Each observer performed one block of 80 trials per condition. The conditions were presented in the following order:

Baseline condition.

- Condition 1 (Movie 7): Three black disks moved left-and-right in concert. Only the middle disk contained a rotating dot. The observers were asked to report the direction of the (non-retinotopic) rotation of the middle disk (clockwise vs. counter-clockwise).
- *Ternus-Pikler display with retinotopic and non-retinotopic rotation*

- Condition 2 (Movie 4): In Condition 2, three disks were presented and the observers were instructed to report the non-retinotopic rotation direction.
- Condition 3 (Movie 4): Condition 3 was identical to Condition 2, but the observers were instructed to *ignore* the non-retinotopic rotation and report only the retinotopic rotation.
- Condition 4 (Movie 3): The observers were again instructed to report the retinotopic rotation direction, but only two disks were presented.
- *Ternus-Pikler display with retinotopic linear motion and non-retinotopic rotation*
- Condition 5 (Movie 2): In Condition 5, three disks were presented and the observers were instructed to report the non-retinotopic rotation direction.
- Condition 6 (Movie 2): Condition 6 was identical to Condition 5, but the observers were instructed to *ignore* the non-retinotopic rotation and report whether the retinotopic rotation *in the left disk* is going up-and-down or left-and-right.
- Condition 7 (Movie 1): The observers were again instructed to report the retinotopic rotation direction, but only two disks were presented.

Control condition

- Condition 8 (Movie 7): As a final control condition, three disks with rotating dot in the middle disk were presented. The dot positions in the left and right disk were each chosen randomly from the 3, 6, 9, 12 o' clock positions and the center of the disk.

Analysis. We analyzed performance in terms of percent correct rotation direction discriminations. We performed pre-planned two-sided paired-samples *t* tests to compare conditions with each other, and two-sided one-sample *t* tests for comparisons with 50% chance-level. All tests were performed in the open-source JASP software (http://jasp-stats.org).

Fixation control. Fixation was automatically controlled after each trial. Fixation was considered broken if the gaze deviated more than 1.5° from the fixation point during a period longer than 20 ms, and when the signal was lost during a period of 150 ms or longer. When fixation was broken, a feedback tone was played and a message was displayed, reminding the participant to fixate the fixation point. Trials with fixation errors were discarded and repeated in the same block at a random later moment.

Results and discussion

In the baseline condition (Condition 1), observers discriminated the rotation direction very well (C1: 84.0%; Figure 2). In Conditions 2-4, we used the stimulus with retinotopic *and* non-retinotopic rotations (Figure 1b). Phenomenologically, only the non-retinotopic rotation was perceived and the retinotopic rotation was largely invisible. In Condition 2, observers reported the non-retinotopic dot rotation in the middle disk. Overall performance was significantly lower than in the baseline condition ($C2_{all}$: 73.3%; $C2_{all}$ vs. C1: $t(9) = 4.40$, $p = .002$, mean diff. = 10.7%). This difference was driven by trials in which the retinotopic rotation direction was opposite of the non-retinotopic rotation direction ($C2_{opposite}$): Performance was strongly deteriorated ($C2_{opposite}$ vs. C1: $t(9) = 6.66$, $p < .001$, mean diff. = 24.7%; $C2_{opposite}$ vs. $C2_{same}$: $t(9) = 6.27$, $p < .001$, mean diff. = 28.0%) and barely exceeded the 50% chance-level (59.3%; $C2_{opposite}$ vs. 50%: $t(9) = 1.90$, $p = .090$). When the retinotopic rotation direction matched the non-retinotopic rotation direction, performance was equally good as in the baseline condition, but not significantly better (87.3%; $C2_{same}$ vs. C1: $t(9) = 1.14$, $p = .284$, mean diff. = 3.3%). A control experiment showed that this is not linked to a ceiling effect (see Experiment 2b). Therefore, it appears that congruent retinotopic rotation has little to no effect on the non-retinotopic motion percept.

In Condition 3, the observers were asked to report the retinotopic rotation direction. Performance was very low, albeit significantly above chance-level ($C3_{all}$: 57.5%; $C3_{all}$ vs. 50%: $t(9) = 3.84$, $p = .004$). In trials where the retinotopic rotation direction matched the visible non-retinotopic rotation direction, performance was reasonably good ($C3_{same}$: 68.2%). However, when the retinotopic and non-retinotopic rotation direction did not match, performance was at chance-level ($C3_{opposite}$: 46.8%; $C3_{opposite}$ vs. 50%: $t(9) = 0.86$, $p < .410$). This pattern of results is compatible with the interpretation that the retinotopic rotation was phenomenologically invisible and that, given the absence of a clear retinotopic percept, the responses were biased by the non-retinotopic rotation. As we saw in Condition 2, the non-retinotopic rotation is clearly perceived when combined with a retinotopic rotation in the same sense, but not when combined with an incongruent retinotopic rotation. This might explain why the responses in Condition 3 were more biased by the (clearly visible) congruent non-retinotopic rotation than the (almost invisible) incongruent non-retinotopic rotation. This suggestion is difficult to test in the case of congruent rotation, where performance was well above chance-level, which could also indicate that the retinotopic rotation was visible. However, the rotation direction of the incongruent retinotopic rotation could *not* be reported with higher than chance accuracy, showing that it was effectively invisible. The strong effect of incongruent retinotopic rotation on the non-retinotopic percept in Condition 2 must hence have been due to *unconscious* processing of the retinotopic rotation. Moreover, the result is not caused by low-level differences between the stimuli. First, the stimuli are identical except for the addition or omission of the third disk. Second, control analyses showed that low-level attributes, such as the initial positions of the disks (left/right), the initial orientation of the rotation (3, 6, 9, 12 o'clock), and the rotation directions (clockwise/counter-clockwise), did not impact performance - the main determinant of performance was whether the retinotopic and non-retinotopic rotations spun in the same or opposite sense (see online supplementary materials).

In Condition 4, we presented only two disks. Now, phenomenologically, the retinotopic, but not the non-retinotopic rotation was perceived. Observers reported the retinotopic rotation direction. Performance was very good, irrespective of whether the now invisible non-retinotopic rotation direction was in the same direction or not ($C4_{all}$: 94.4%, $C4_{same}$: 95.8%, $C4_{opposite}$: 93.0%; $C4_{same}$ vs. $C4_{opposite}$: $t(9) = 2.18$, $p = .057$, mean diff. = 2.8%).

In Conditions 5-7, we presented the Ternus-Pikler display with linear retinotopic dot motions (Figure 1a). In Condition 5, where we presented three disks, only the non-retinotopic dot rotation in the middle disk was perceived, whereas the retinotopic up-down dot motion in one, and left-right dot motion in the other disk were invisible. Observers were asked to report the non-retinotopic rotation direction and performed equally well as in the baseline condition without dots in the left and right disk (C5: 84.9%; C5 vs. C1: $t(9) =0.43$, $p = .678$, mean diff. = 0.9%) and as in the condition where the retinotopic rotation direction matched the non-retinotopic direction (C5 vs. $C2_{same}$: $t(9) =0.95$, $p = .367$, mean diff. = 2.4%). Hence, linear retinotopic motion did not affect the non-retinotopic motion percept. In Condition 6, we presented three disks and asked the observers to report the invisible retinotopic motion direction (up-down vs. left-right). Performance was at chance-level (C6: 54.7%; C6 vs. 50%: $t(9) = 0.98$, $p = .352$), confirming that the retinotopic motion was indeed invisible. In Condition 7, we presented only two disks, and the retinotopic rotation could be readily reported (C7: 92.9%).

Finally, in Condition 8, we presented a Ternus-Pikler display with a non-retinotopic rotation in the middle disk. The dot positions in the left and right disk were chosen at random, creating random motion percepts in these disks. Observers reported the non-retinotopic rotation direction. Performance was good (78.7%) and not significantly different from the baseline condition (C8 vs. C1: $t(9) =1.67$, $p = .130$, mean diff.

= 5.3%), although the random dot motions in the left and right disks increased the complexity of the stimulus. However, performance in Condition 8 was slightly lower than when the retinotopic and non-retinotopic rotations were in the same direction (C8 vs. $C2_{same}$: $t(9) = 2.77$, $p = .022$, mean diff. = 8.6%).

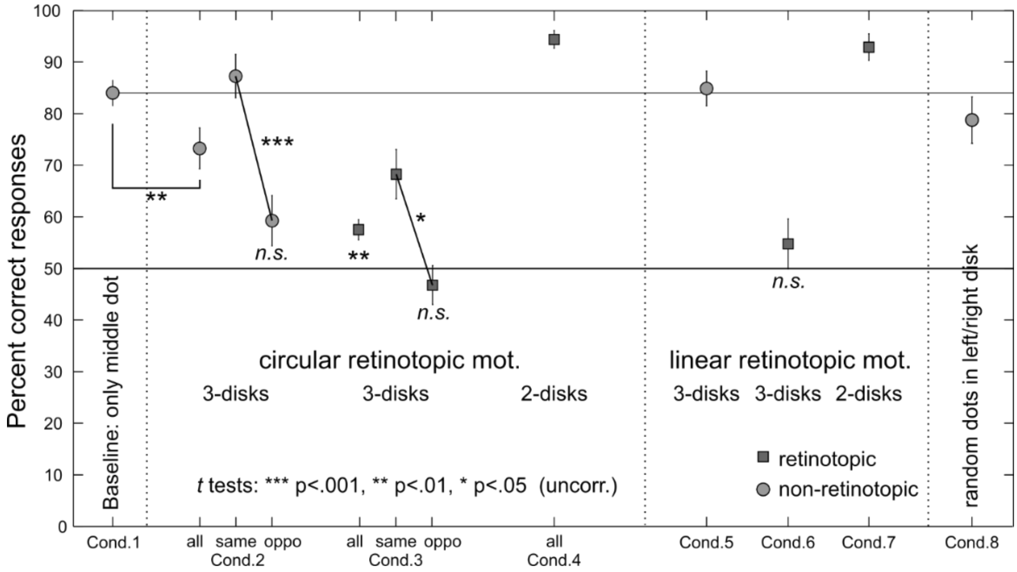


Figure 2. Motion direction discrimination performance in Experiment 1a. When three disks are presented, the non-retinotopic motion is perceived (C1, C2, C5, C8). The non-retinotopic motion percept is impaired when the stimulus with circular retinotopic motion is used (C2). The impairment occurs because performance is not significantly different from chance-level when the retinotopic rotation direction is incompatible with the non-retinotopic rotation direction ($C2_{opposite}$). This is highly surprising, because the retinotopic motion is largely invisible: When three disks are presented, the non-retinotopic motion percept dominates the percept (C3, C6). The retinotopic motion is visible when only two disks are presented (C4, C7). Error bars depict one standard error of the mean.

Experiment 1b: The visibility of non-retinotopic motion when two disks are presented

In Experiment 1a, we found that the direction of an invisible (i.e. unreportable) retinotopic rotation affects conscious non-retinotopic rotation perception. When we presented only two disks, the retinotopic rotation was well perceived irrespective of the non-retinotopic rotation direction. We did not test whether the non-retinotopic rotation was invisible in this case. We therefore performed a second experiment, in which we included a condition with two disks in which we asked the observers to report the non-retinotopic rotation direction.

Methods

The methods for Experiment 1b were identical to Experiment 1a unless noted otherwise below. Ten new, naïve observers participated (mean age 23.6 years, *SD* = 3.0 years, all male, 2 left-handed, 1 wore glasses, no exclusions). Stimuli were presented on a 24inch Asus VG24248QE LCD monitor (1920x1080 pixels, 60 Hz, http://display-corner.epfl.ch).

Conditions 1-3 were identical to Conditions 2-4 of Experiment 1a. Conditions 5-7 were identical to Conditions 5-7 of Experiment 1a. In Condition 4, we presented only two disks and asked our observers to report the non-retinotopic rotation. Each observer performed one block of 80 trials per condition, except for

Condition 2 (2 blocks of 80 trials) and Conditions 3 and 7 (each 1 block of 48 trials). The conditions were presented in order.

Results and discussion

We first presented the Ternus-Pikler display with three disks and invisible retinotopic rotation. In Condition 1, observers reported the non-retinotopic rotation direction. Overall performance was good ($C1_{all}$: 79.2%; Figure 3). As in Experiment 1a, performance was strongly impaired when the retinotopic rotation was in the opposite direction, compared to when it was in the same direction as the non-retinotopic rotation ($C1_{same}$: 89.5%, $C1_{opposite}$: 69.0%; $C1_{same}$ vs. $C1_{opposite}$: $t(9) = 6.49$, $p < .001$, mean diff. = 20.5%). Again performance levels were comparable when the retinotopic rotation direction was the same as for the non-retinotopic rotation and when a linear retinotopic motion was used (C6: 85.6%; C6 vs. $C1_{same}$: $t(9) = 1.62$, $p = .140$, mean diff. = 3.9%).

In Condition 3, we presented three disks and observers reported the retinotopic rotation direction. Performance was very low overall ($C2_{all}$: 57.5%), albeit significantly above chance level ($C2_{all}$ vs. 50%: $t(9) = 3.17$, $p = .011$). Performance was significantly better when the retinotopic and non-retinotopic rotation direction matched than when they mismatched ($C2_{same}$: 73.6%, $C2_{opposite}$: 41.4%; $t(9) = 5.12$, $p < .001$, mean diff. = 32.3%). Performance in the opposite rotation case was even significantly *below* chance-level ($t(9) = 2.85$, $p = .019$). As in Experiment 1a, this pattern of results is compatible with the interpretation that the retinotopic rotation is invisible and observers' responses are biased by the non-retinotopic rotation.

In Condition 3 and 4, only two disks were presented and the retinotopic, but not the non-retinotopic rotation was perceived. In Condition 3, observers reported the retinotopic rotation direction and performance was very good ($C3_{all}$: 92.9%), irrespective of whether the non-retinotopic rotation was in the same or the opposite direction ($C3_{same}$: 93.3%, $C3_{opposite}$: 92.5%; $t(9) = 0.36$, $p = .726$). In Condition 4, the observers were asked to report the (invisible) non-retinotopic rotation direction. Similar to the pattern observed in Condition 2, performance was reasonably good when the retinotopic and non-retinotopic rotation were in the same direction ($C4_{same}$: 74%) and below chance when not ($C4_{opposite}$: 34.5%), indicating that the non-retinotopic rotation was invisible and observers' responses were biased by the visible retinotopic rotation.

In Conditions 5-7, we used the Ternus-Pikler display with linear retinotopic motion. As in Experiment 1a, non-retinotopic rotation discrimination performance was good and comparable to trials of Condition 1, where the retinotopic and non-retinotopic rotations were in the same direction (C5: 85.6%). The retinotopic motion direction (up-down vs. left-right) could *not* be reported above chance-level when three disks were presented (C6: 55.7%; C6 vs. 50%: $t(9) = 1.31$, $p = .222$). The retinotopic motion could easily be reported when only two disks were presented (C7: 88.1%).

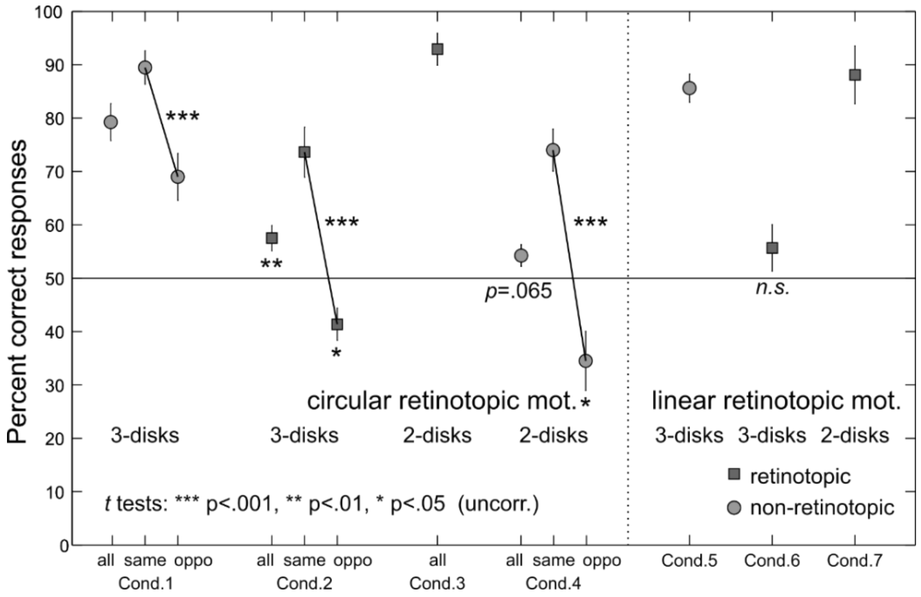


Figure 3. Motion direction discrimination performance in Experiment 1b. In line with the Experiment 1a, when three disks are presented the non-retinotopic motion is perceived (C1, C5) and the retinotopic motion cannot reliably be reported (C2, C6). When only two disks are presented, the retinotopic motion is perceived (C3, C7) and non-retinotopic motion cannot reliably be reported (C4). When the stimulus with retinotopic and non-retinotopic rotation is used, observers tend to answer in line with the visible rotation even if it is not task relevant, leading to above chance-level performance when the rotation directions are in the same direction ($C2_{same}$, $C4_{same}$) and below chance-level performance when not ($C2_{opposite}$, $C4_{opposite}$). Replicating the first experiment, the non-retinotopic motion direction can equally well be reported then the retinotopic rotation direction matches the non-retinotopic rotation direction ($C1_{same}$) and when the retinotopic motion is linear, rather than circular (C5). However, the non-retinotopic *motion percept is strongly degraded if the retinotopic rotation is in the opposite direction. Error bars show one standard error of the mean.*

Experiment 2a: Influence of multiple retinotopic and non-retinotopic rotations

In Experiment 1, we found that conscious non-retinotopic rotation perception is impaired by unconscious processing of a retinotopic rotation in the opposite direction. Linear retinotopic motion and a congruent retinotopic rotation did not affect the percept. In contrast, when we presented only two disks, retinotopic motion was perceived and was unaffected by the invisible non-retinotopic rotation.

In Experiment 2, we tested whether multiple retinotopic rotations interact. We added a condition in which there was not one, but two retinotopic rotations (Figure 4a and Movie 9). We were primarily interested if one congruent and one incongruent retinotopic rotation would cancel each other, and whether the detrimental effects of two incongruent retinotopic rotations would add up, leading to even worse performance. In addition, we investigated whether visible *non*-retinotopic rotations affect the percept in a similar fashion as invisible retinotopic rotations. To this end, we added a condition in which we presented non-retinotopic rotation in each of the three disks, but no retinotopic rotations (Figure 4b and Movie 13).

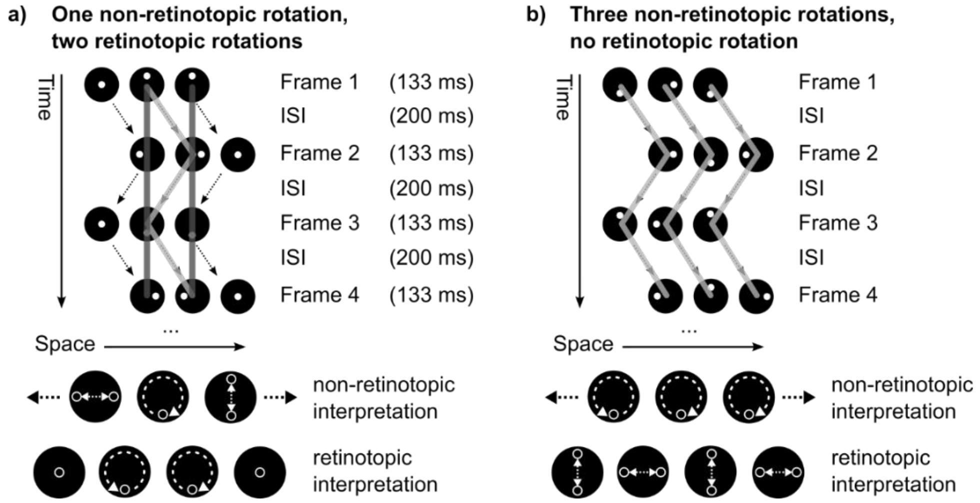

a) One non-retinotopic rotation, two retinotopic rotations
Time
Frame 1 (133 ms)
ISI (200 ms)
Frame 2 (133 ms)
ISI (200 ms)
Frame 3 (133 ms)
ISI (200 ms)
Frame 4 (133 ms)
...
Space
non-retinotopic interpretation
retinotopic interpretation
b) Three non-retinotopic rotations, no retinotopic rotation
Time
Frame 1
ISI
Frame 2
ISI
Frame 3
ISI
Frame 4
...
Space
non-retinotopic interpretation
retinotopic interpretation

Figure 4. a) In Experiment 2 we added a condition in which we presented two retinotopic rotations, instead of one as in Experiment 1. The two retinotopic rotations could each spin in the same or opposite direction of the non-retinotopic rotation in the middle disk. In the depicted example, the non-retinotopic rotation (light grey) and right retinotopic rotation (dark grey) are clockwise, and the left retinotopic rotation is counter-clockwise. Only the non-retinotopic interpretation is perceived: The dot in the middle disk rotates and the dots in the left and right disk move up-and-down and left-and-right, respectively. The retinotopic interpretation of the image was invisible and is shown only to illustrate the two retinotopic rotations. b) We also added a condition in which no retinotopic, but three non-retinotopic rotations were presented. The observers reported the rotation in the middle disk. The outer disks could rotate both in the same direction as the middle disk, both in the opposite direction, or one in the same and one in the opposite direction. Retinotopically, the dots moved left-right or up-down every second frame. In the depicted example, the dot rotates clockwise in the middle and right disk and counter-clockwise in the left disk. Again, only the non-retinotopic interpretation was perceived. The retinotopic interpretation was invisible and is shown only for illustration.

Methods

The methods for Experiment 2a were identical to Experiment 1a unless noted otherwise below. Ten new, naïve observers participated (mean age 24.4 years, *SD* = 1.6 years, 6 female, 2 left-handed, 1 with glasses, no exclusions).

Conditions 1-3 were identical to Conditions 1-3 of Experiment 1a. In Condition 1, we determined the baseline performance level using a Ternus-Pikler display with non-retinotopic rotation in the middle disk, and no dots in the left and right disk (Movie 7). In Conditions 2 and 3, we used the Ternus-Pikler display with three disks and one retinotopic and one non-retinotopic rotation (cf. Experiment 1, Figure 1b, and Movie 4). The observers reported the non-retinotopic rotation direction in Condition 2 and the retinotopic rotation in Condition 3. In Condition 4, there were retinotopic rotations in *two* disks, and the dot in the third disk was always in the center (Figure 4a and Movie 9). The directions of both retinotopic and the non-retinotopic rotations were combined in a factorial design and presented in random order. Hence, either both retinotopic rotations spun in the opposite direction of the non-retinotopic rotation (Movie 10), both in the same direction (Movie 11), or one in the same and one in the opposite direction (Movie 12). In Condition 5, we presented visible *non-retinotopic* rotations in all three disks and observers reported the non-retinotopic rotation direction of the middle disk (Figure 4b and Movie 13). The left and right disk rotation could either both spin in the opposite sense of the middle disk rotation (Movie 14), both in the same sense (Movie 15), or one in the same and one in the opposite sense (Movie 16). There was no retinotopic rotation. Retinotopically the dots moved linearly up-down or left-right in every second frame. In Condition 6, there was a non-retinotopic rotation in the middle disk and the dots in the left and right disk were positioned randomly, with the restriction that only linear motion could occur between the randomly chosen dot positions. This was done to control whether the trend towards decreased performance in a similar condition in Experiment 1 can be explained by an incidental occurrence of incongruent rotation in the randomly placed dots. Whether the stimulus was first presented on the left or right and the initial orientation of the non-retinotopic rotation in the middle disk was chosen randomly with equal probability.

Observers performed one block of 80 trials per condition. Other than in Experiment 1a, all conditions were run in random order and with three disks. Before the experiment, each observer performed training blocks with the baseline stimulus of 20 trials each, until performance levels were above 70%. Seven observers met this criterion already in the first training block, two observers in the second training block, and one observer in the third training block.

Results and discussion

In Experiment 2a, Conditions 1-3 were identical to Conditions 1-3 of Experiment 1a and showed the same effects (Figure 5). In Condition 1, observers reported the direction of a non-retinotopic dot rotation in the middle disk and all other dots were omitted. Performance was very good (C1: 91.1%). In Conditions 2-3, we used the Ternus-Pikler display with one invisible retinotopic rotation (Figure 1b). In Condition 2, observers reported the non-retinotopic rotation direction. Performance was lower than in Condition 1 ($C2_{all}$: 83.9%; $C2_{all}$ vs. C1: *t*(9) = 2.16, *p* = .059, mean diff. = 7.3%), which was again mainly due to lower performance in trials in which the retinotopic and non-retinotopic rotation direction were in opposite directions ($C2_{opposite}$: 77.0%; $C2_{opposite}$ vs. $C2_{same}$: *t*(9) = 2.30, *p* = .047, mean diff. = 13.8%). When the retinotopic and non-retinotopic rotation were in the same direction, performance was equally good as in Condition 1 ($C2_{same}$: 90.7%; $C2_{same}$ vs. C1: *t*(9) = 0.13, *p* = .898, mean diff. = 0.4%). As previously mentioned, a control experiment showed that this was not due to a ceiling effect (see Experiment 2b). In Condition 3, observers were asked to report the invisible retinotopic rotation direction and performance was hardly above chance-level ($C3_{all}$: 57.5%; $C3_{all}$ vs. 50%: *t*(9) = 2.26, *p* = .050). Performance was at chance-level when the direction of the retinotopic and non-retinotopic rotation were incongruent ($C3_{opposite}$: 49.2%; $C3_{opposite}$ vs. 50%: *t*(9) = 0.14, *p* = .891), and moderately good when they were congruent ($C3_{same}$: 65.7%; $C3_{opposite}$ vs. 50%: *t*(9) = 4.10, *p* = .003).

In Condition 4, we added a second invisible retinotopic rotation to the stimulus (Figure 4a). Observers reported the non-retinotopic rotation direction. Overall performance was worse than in Condition 1 ($C4_{all}$: 79.9%; $C4_{all}$ vs. C1: *t*(9) = 3.94, *p* = .003, mean diff. = 11.3%). When both retinotopic rotations were in the same direction as the non-retinotopic rotation, performance was equally good as in Condition 1 ($C4_{same}$: 93.0%; $C4_{same}$ vs. C1: *t*(9) = 0.74, *p* = .481, mean diff. = 1.9%). However, when one of the two retinotopic rotations was in the opposite direction and the other in the same direction of the non-retinotopic rotation, performance was lower compared to when both were in the same direction ($C4_{same/opposite}$: 81.1%; $C4_{same/opposite}$ vs. $C4_{same}$: *t*(9) = 4.16, *p* = .002, mean diff. = 11.9%) and compared to Condition 1 ($C4_{same/opposite}$ vs. C1: *t*(9) = 3.49, *p* = .007, mean diff. = 10.0%). When both retinotopic rotations were in the opposite direction of the non-retinotopic rotation, performance was even lower ($C4_{opposite}$: 64.2%; $C4_{opposite}$ vs. $C4_{same/opposite}$: *t*(9) = 3.67, *p* = .005, mean diff. = 16.9%). Hence, only the negative effect of the opposite direction retinotopic rotations added up, but one same and one opposite direction rotation did not cancel each other.

In Condition 5, we presented non-retinotopic rotations in all three disks, but no retinotopic rotation (Figure 4b). All three rotations were perceived and observers reported the non-retinotopic rotation in the middle disk. Overall, performance in this condition was slightly worse than in Condition 1 ($C5_{all}$: 81.0%; $C5_{all}$ vs. C1: *t*(9) = 3.42, *p* = .008, mean diff. = 10.1%). But performance differed only very slightly depending on whether the non-retinotopic rotations in the flanking disks were both in the same ($C5_{same}$: 83.5%) or both in the opposite direction of the middle disk ($C5_{opposite}$: 77.0%; $C5_{same}$ vs. $C5_{opposite}$: *t*(9) = 1.86, *p* = .096, mean diff. = 6.5%). When one of the rotations was in the same and one in the opposite direction of the middle disk rotation, performance did not differ significantly compared to when both were in the same or in the opposite direction (C5$_{\text{same/opposite}}$: 81.8%; C5$_{\text{same/opposite}}$ vs. C5$_{\text{same}}$: $t(9) = 0.46$, $p = .655$, mean diff. = 1.7%; C5$_{\text{same/opposite}}$ vs. C5$_{\text{opposite}}$: $t(9) = 1.98$, $p = .079$, mean diff. = 4.8%).

Finally, in Condition 6 we presented the non-retinotopic rotation in the middle disk with randomly placed dots in the left and right disks. Performance was slightly lower than in Condition 1, where the left and right disk had no dots at all (C6: 83.1%; C6 vs. C1: $t(9) = 2.79$, $p = .021$, mean diff. = 8.0%). This is somewhat surprising, since performance did not decrease significantly in a similar condition of Experiment 1a, where we did not prevent the incidental occurrence of rotations in the random dots.

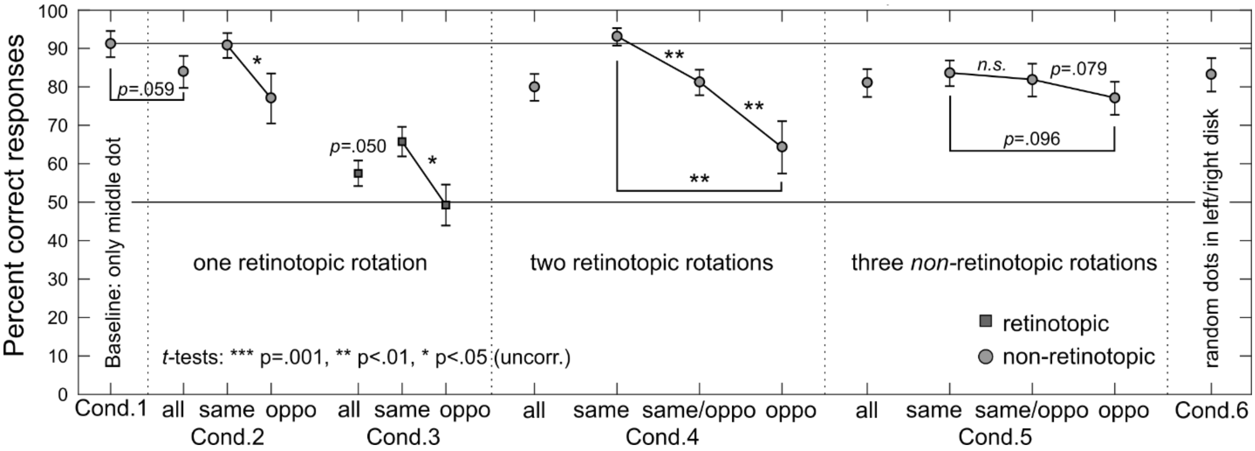


Figure 5. Motion direction discrimination performance in Experiment 2a. Conditions 1-3 replicate the findings of Experiment 1: Unconscious processing of an incongruent retinotopic rotation interferes with non-retinotopic rotation perception. In Condition 4, we used a stimulus with two retinotopic rotations. Performance decreased when one retinotopic rotation direction was incongruent with the non-retinotopic rotation, compared to when all were in the same direction. Performance decreased even more when both retinotopic rotations were incongruent with the non-retinotopic rotation. In Condition 5, we presented non-retinotopic rotations in all three disks and there was no retinotopic rotation. The participants reported the rotation in the middle disk. Performance was slightly lower than in the baseline condition (Condition 1), and differed only little depending on whether the left and right disks rotated in the same or opposite sense of the middle disk. In Condition 6, the dots in the left and right disk were placed randomly and the middle disk rotated. Performance was slightly lower than in the baseline condition. Error bars depict one standard error of the mean.

Experiment 2b: Control for ceiling effects

In Experiments 1a-b and 2a, we found that incongruent invisible retinotopic motion degraded the conscious non-retinotopic motion percept, but congruent retinotopic motion did not improve it. However, mean performance was very good in the baseline condition of these experiments (84-91% correct), so that the range of possible improvement was very limited (possibility of a ceiling effect). In Experiment 2b, we individually adjusted the stimulus eccentricity, stimulus duration, and ISI duration to increase task difficulty and achieve a lower mean baseline performance of only 78.5 %. We thereby avoided a ceiling effect and could test whether non-retinotopic rotation perception is facilitated by invisible retinotopic rotation spinning in the same direction.

Methods

The conditions of Experiment 2b were identical to Conditions 1, 2, and 4 of Experiment 2a. Prior to the experiment, stimulus eccentricity (4-8° above fixation), stimulus duration (83-133 ms), and ISI duration (133-200 ms) were individually adjusted for each observer to modify task difficulty and prevent ceiling performance. Nineteen new, naïve observers participated. Nine observers whose performance in the baseline condition was not in the range of 65-85 % correct responses were excluded from analysis. Hence, ten observers were available for analysis (mean age 22.6 years, *SD* = 4.7 years, 4 female, all right-handed, 1 with glasses).

Results and discussion

As intended, mean performance in the baseline condition was lower than in the previous experiments ($C1_{all}$ = 78.5 %; Figure 6). Nevertheless, we did not find evidence that perception of the non-retinotopic dot rotation is facilitated by retinotopic rotations spinning in the same sense. With *one* congruent retinotopic rotation performance was even significantly lower than in the baseline condition ($C2_{same}$: 69.5 %; $C1_{all}$ vs. $C2_{same}$: $t(9) = 2.68$, $p = .025$, mean diff. = 9.0 %). With *two* congruent rotations, performance was not significantly better than in the baseline condition ($C3_{same}$: 82.3 %; $C1_{all}$ vs. $C3_{same}$: $t(9) = 1.23$, $p = .248$, mean diff. = 3.8 %).

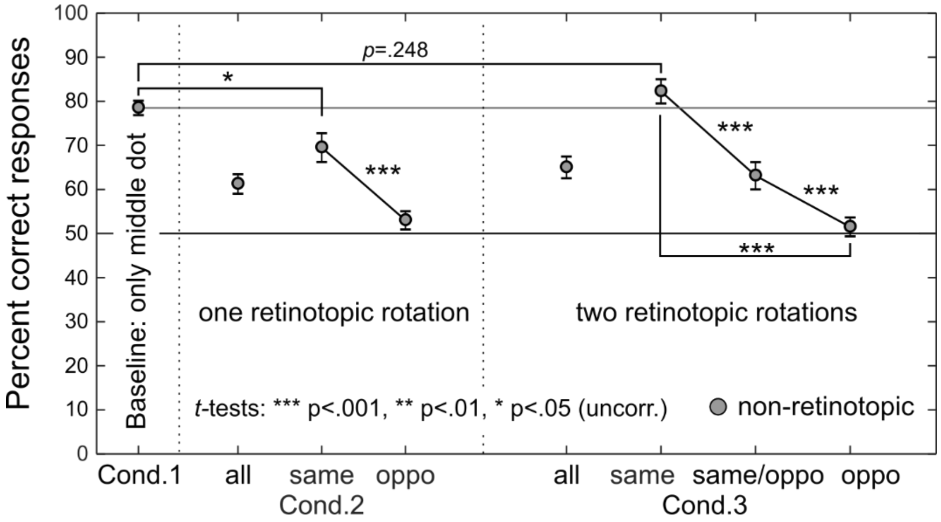


Figure 6. Motion direction discrimination performance in Experiment 2b. To avoid ceiling effects, stimulus eccentricity, stimulus duration, and ISI duration were individually adjusted before the experiment. Observers reported the non-retinotopic rotation direction. As in the previous experiments, performance was impaired when one incongruent retinotopic rotations was present (C2oppo, C3same/oppo). Performance decreased even more when two, instead of one, incongruent retinotopic rotations were presented (C3oppo). However, the presence of one (C2same) or even two (C3same) congruent retinotopic rotations did not improve performance.

General Discussion

Research on unconscious perception traditionally involves the presentation of a stimulus that is visible when presented alone, but invisible when presented together with a second stimulus (Bachmann et al., 2007). Explanations of invisibility are ex- or implicitly based on retinotopic inhibitory circuits. For example in masking, the processing of the first stimulus is inhibited by the second stimulus in early, retinotopic visual areas (e.g., Breitmeyer & Ganz, 1976; Enns & DiLollo, 2000; Lamme, 2006; Fahrenfort, Scholte, & Lamme, 2007; Breitmeyer & Öğmen, 2006; Noory et al., 2015). The invisible stimulus, even though it is not consciously perceived, can alter subsequent processes. For example, in masked priming, an invisible prime can speed up or delay responses to subsequently presented targets. This effect is typically explained by a short-lived preactivation of the motor system (e.g., Klotz & Wolff, 1995; Klotz & Neumann, 1999).

Invisibility occurs not only in artificial experimental settings, but is a fundamental part of everyday vision (Herzog et al., 2014). For example, the organization of stimulus parts into a Gestalt (“object”) results in a non-retinotopic reference-frame that renders the *retinotopic* motion of object parts invisible, while revealing *non-retinotopic* motions that are computed relative to the non-retinotopic reference-frame (Öğmen & Herzog, 2010). For example, a bicycle reflector is perceived to circle, because it is perceived relative to the moving bicycle, and its cycloidal retinotopic trajectory is invisible (cf. Johansson, 1950; Duncker, 1929). Here, we asked whether unconscious processing of retinotopic information affects conscious non-retinotopic perception. We used a modified Ternus-Pikler display, in which dot motion is perceived non-retinotopically and the retinotopic dot-motion is invisible. We found that unconscious retinotopic processing can influence conscious non-retinotopic processing to a substantial degree. As in ambiguous figures, all display elements were fully visible and only the retinotopic interpretation(s) of the display were suppressed. Nonetheless, the suppressed retinotopic motion interpretation strongly affected the conscious non-retinotopic motion percept. Similar effects have been reported with ambiguous figures: Conscious perception of the background was suppressed. Nonetheless, unconscious processing of the background caused responses to objects presented later to be slower when they were semantically related, compared to when they were unrelated (Peterson & Kim, 2001; Peterson & Skow, 2008).

The mechanisms that are thought to explain invisibility in masking or ambiguous figures cannot account for invisibility of the retinotopic motion in our paradigm. In masking and similar techniques, the stimulus itself is thought to be invisible because its processing in early retinotopic areas of visual cortex is suppressed. In ambiguous figures, the stimulus itself remains visible and only an interpretation of the stimulus is suppressed, as in our paradigm. But also in the case of ambiguous figures the effect can be explained by competition between retinotopically organized boundary ownership neurons (Zhou, Friedman, & von der Heydt, 2000; Zhaoping, 2005; Layton, Mingolla, & Yazdanbakhsh, 2012). Obviously, retinotopic inhibition cannot explain the invisibility of the retinotopic interpretations in our paradigm, because the disks and dots themselves are clearly visible. The retinotopic motion must be suppressed on an object level, rather than a retinotopic level. The retinotopic motion is only invisible when the disks are perceived to move. This motion percept depends on complex spatio-temporal processing, which determines the grouping of the disk. Most importantly, the effect is largely independent of the stimulus layout in retinotopic coordinates. For example, one can use larger gaps between the disks, present the disks at different relative positions, or use different elements than disks, such as squares or diamonds (Boi et al., 2009; Lauffs, Öğmen, & Herzog, 2017; Petersik & Rice, 2006).

Finally, in masking the strong effects of the unconscious prime on conscious elements are usually explained by a pre-activation of the motor system. In congruent cases, reaction times speed up because the “correct” response is pre-activated, in incongruent cases they slow down for the same reason. For these effects to occur, the target has to be presented rapidly after the prime, typically not more than 250 ms later (e.g., Jacob, Breitmeyer, & Treviño, 2013). Such motor priming cannot explain our results. First, at least two stimulus frames of 133 ms, interspersed by a 200 ms ISI (altogether 466 ms), are needed to perceive dot motion. Second, reaction times play no role in our paradigm because responses are delayed until after the last stimulus frame (without dot) had disappeared, which happened 333 ms after the last frame with dot had offset.

What mechanism can explain the influence of the invisible retinotopic rotation on the conscious non-retinotopic percept? By adding a second retinotopic rotation to the Ternus-Pikler display, we could investigate how multiple invisible retinotopic rotations interact during unconscious processing. In the simplest model that comes to mind, unconscious retinotopic motion signals are summed up at a retinotopic integration stage, which then influences the non-retinotopic, conscious percept. Indeed, when two retinotopic rotations are incongruent with respect to the non-retinotopic rotation, performance changes much more strongly than when only one incongruent motion is presented (Figure 5, Condition 4). However, this model fails for two reasons. First, one incongruent and one congruent retinotopic rotation do not cancel each other. Performance decreases similarly to the condition with only one incongruent and no congruent rotation (Figure 5, compare Condition 4 and 2). Hence, the congruent rotation seems not to matter. Second, congruent retinotopic rotations do not influence the non-retinotopic motion percept (Figure 6, Condition 2). This holds true even when *two* congruent rotations are presented (Figure 6, Condition 3). In this case the two retinotopic and the non-retinotopic rotations spin in the same direction. Also linear retinotopic motion had no influence on the non-retinotopic percept, further indicating that it is not the sheer retinotopic processing load that changes performance. Hence, it seems that inhibition of the non-retinotopic percept is caused selectively by *incongruent* retinotopic information.

Our results may be summarized by a model in which retinotopic rotations that spin in the same direction (e.g., all clockwise) are integrated and inhibit perception of incongruent non-retinotopic rotation, but do not affect perception of congruent non-retinotopic rotation (Figure 7). The effects are substantial. In the condition with two incongruent retinotopic rotations, performance dropped by about 20-30 % compared to a no motion condition and the congruent and linear motion conditions (Figure 5, Condition 4).

In summary, the Ternus-Pikler display is a versatile tool to investigate long-lasting unconscious processing without suppressing the elements of the display itself. Traditional retinotopic-mechanisms are unlikely at work since what is perceived unconsciously and consciously depends on Gestalt formation through complex spatio-temporal processing, such as establishing group versus element motion. The Ternus-Pikler display allows to investigate how unconscious, retinotopic processing interacts with each other and the conscious non-retinotopic processing, and how rivaling interpretations influence each other. The Ternus-Pikler display can be flexibly adjusted to the needs of the researcher, for example by adding additional disks and rotations, or using other visual features (see e.g., Boi et al., 2009). Thereby, it allows to ask different questions that cannot be answered with traditional paradigms.

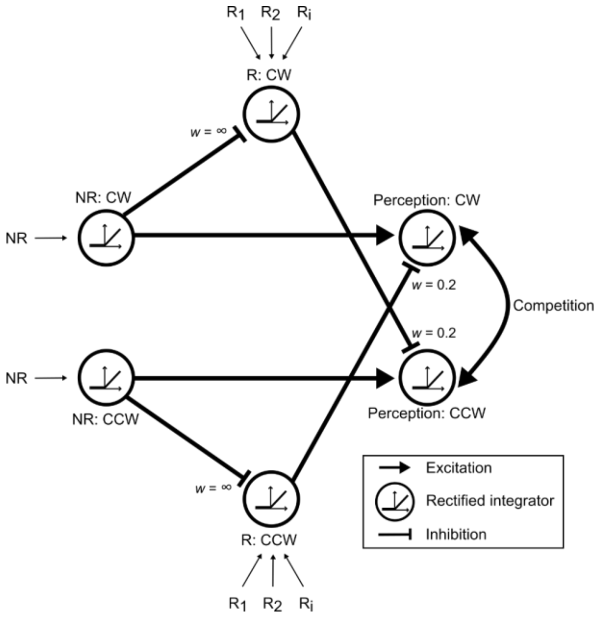


Figure 7. A simple model that summarizes our findings. We found that interactions between retinotopic and non-retinotopic motion signals are specific to rotation-type. Accordingly, the model considers interactions between neurons tuned to rotational motion exclusively (linear and random motion computations are not included). Furthermore, non-retinotopic motion is perceived only when a non-retinotopic reference-frame is established in the group-motion conditions (i.e., the conditions with three disks). Hence, the model focuses on the conditions with three disks.

Clockwise (CW) and counter-clockwise (CCW) retinotopic (R) and non-retinotopic (NR) rotations are detected by motion-sensitive neurons. Each neuron is only sensitive to one rotation direction. Retinotopic CW and CCW signals are integrated separately, using a rectifier as transfer function. That is, the neuron activates only for the rotation direction it is sensitive to, and the output-strength depends on the input-strength. “Non-retinotopic CW neurons” project to neurons coding for the CW percept (“perception neurons”) and fully inhibit (w= INFINITY) the “retinotopic CW integration neurons”, and likewise for the “CCW neurons”. Because of the inhibition, congruent retinotopic motion does not contribute to the overall percept. The retinotopic CW neurons inhibit the CCW perception neurons and the retinotopic CCW neurons inhibit the CW perception neurons. This inhibition accounts for the deterioration in non-retinotopic motion perception by incongruent retinotopic signals. Competition between the perception neurons determines the final percept. As an example, if multiple retinotopic rotations (R_1, R_2, … R_i) are integrated at the detector stage, the perception neuron for incongruent non-retinotopic rotation receives stronger inhibition. However, this inhibition is weighted with a low factor (here arbitrarily chosen as w=0.2), so that retinotopic rotations weaken, but typically do not fully suppress the perception of incongruent non-retinotopic rotation.

The model reproduces our experimental results well: 1) Congruent retinotopic rotation neither deteriorates nor improves perception of the non-retinotopic rotation because of the strong inhibition from non-retinotopic cells to their congruent retinotopic counter-parts. 2) Incongruent retinotopic rotation impairs perception of the non-retinotopic rotation because of the inhibition from retinotopic neurons to incongruent perception-neurons. 3) Multiple incongruent retinotopic rotations impair non-retinotopic rotation perception more than a single incongruent retinotopic rotation because of the integration of inputs at retinotopic neurons. 4) One congruent and one incongruent retinotopic rotation do not cancel each other because the congruent retinotopic neuron is itself suppressed.

Acknowledgements

We thank Victoire Gorden for her help with data collection.

Funding

This work was supported by the Swiss National Science Foundation (SNSF), grant 153001 ”Basics of visual processing: From retinotopic encoding to non-retinotopic representations”.

Conflicts of interest

All authors declare to have no conflict of interest.

References

Ağaoğlu, M. N., Clarke, A. M., Herzog, M. H., & Öğmen, H. (2016). Motion-based nearest vector metric for reference frame selection in the perception of motion. *Journal of vision*, *16*(7), 14-14.

Bach, M. (1996). The Freiburg Visual Acuity Test-automatic measurement of visual acuity. *Optometry & Vision Science*, *73*(1), 49-53.

Bachmann, T., Breitmeyer, B. G., Öğmen, H. (2007). *Experimental Phenomena of Consciousness: A Brief Dictionary*. Oxford University Press: New York, N.Y.

Bachmann, T., & Francis, G. (2013). *Visual masking: Studying perception, attention, and consciousness*. Oxford, UK: Academic Press.

Blake, R. (1989). A neural theory of binocular rivalry. *Psychological Review*, *96*(1), 145.

Blake, R. (2001). A primer on binocular rivalry, including current controversies. *Brain and mind*, *2*(1), 5-38.

Boi, M., Öğmen, H., Krummenacher, J., Otto, T. U., & Herzog, M. H. (2009). A (fascinating) litmus test for human retino-vs. non-retinotopic processing. *Journal of Vision*, *9*(13), 5-5.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological review*, *83*(1), 1.

Breitmeyer, B., & Öğmen, H. (2006). *Visual masking: Time slices through conscious and unconscious vision* (No. 41). Oxford University Press.

Clarke, A. M., Öğmen, H., & Herzog, M. H. (2016). A computational model for reference-frame synthesis with applications to motion perception. *Vision Research*, *126*, 242-253.

Duncker, K. (1929). Über induzierte Bewegung. *Psychologische Forschung*, *12*(1), 180–259.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex*, *7*(2), 181-192.

Enns, J. T., & Di Lollo, V. (2000). What’s new in visual masking? *Trends in Cognitive Sciences*, *4*(9), 345-352.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, *19*(9), 1488-1497.

Fehrer, E., & Raab, D. (1962). Reaction time to stimuli masked by metacontrast. *Journal of experimental psychology*, *63*(2), 143.

Hein, E., & Cavanagh, P. (2012). Motion correspondence in the Ternus display shows feature bias in spatiotopic coordinates. *Journal of Vision*, *12*(7), 16-16.

Hein, E., & Moore, C. M. (2012). Spatio-temporal priority revisited: The role of feature identity and similarity for object correspondence in apparent motion. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 975.

Herzog, M. H., Hermens, F., & Öğmen, H. (2014). Invisibility and interpretation. *Frontiers in Psychology*, *5*.

Jacob, J., Breitmeyer, B. G., & Treviño, M. (2013). Tracking the first two seconds: three stages of visual information processing? *Psychonomic bulletin & review*, *20*(6), 1114-1119.

Johansson, G. (1950). Configurations in the perception of velocity. *Acta Psychologica*, *7*, 25-79.

Johansson, G. (1974). Vector analysis in visual perception of rolling motion. *Psychological Research*, *36*(4), 311-319.

Johansson, G. (1976). Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research*, *38*(4), 379-393.

Klotz, W., & Wolff, P. (1995). The effect of a masked stimulus on the response to the masking stimulus. *Psychological research*, *58*(2), 92-101.

Klotz, W., & Neumann, O. (1999). Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 976.

Kovacs, I., Papathomas, T. V., Yang, M., & Fehér, Á. (1996). When the brain changes its mind: interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences*, *93*(26), 15508-15511.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494-501.

Layton, O. W., Mingolla, E., & Yazdanbakhsh, A. (2012). Dynamic coding of border-ownership in visual cortex. *Journal of Vision*, *12*(13), 8-8.

Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature*, *379*(6565), 549.

Lauffs, M. M., Öğmen, H., & Herzog, M. H. (2017). Unpredictability does not hamper nonretinotopic motion perception. *Journal of Vision*, *17*(9), 6-6.

Noory, B., Herzog, M. H., Öğmen, H. (2015). Retinotopy of visual masking and non-retinotopic perception during masking. *Attention, Perception, & Psychophysics*, *77*, 1263-1284.

Öğmen, H. & Herzog, M. H. (2010). The geometry of visual perception: Retinotopic and nonretinotopic representations in the human visual system. *Proceedings of the IEEE, 98*(3), 479-492.

Petersik, J. T., & Rice, C. M. (2006). The evolution of explanations of a perceptual phenomenon: A case history using the Ternus effect. *Perception*, *35*(6), 807-821.

Peterson, M. A., & Kim, J. H. (2001). On what is bound in figures and grounds. *Visual Cognition*, *8*(3-5), 329-348.

Peterson, M. A., & Skow, E. (2008). Inhibitory competition between shape properties in figure-ground perception. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 251.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437-442.

Pikler, J., (1917). *Sinnesphysiologische Untersuchungen*. Leipzig, Germany: Barth.

Ternus, J., (1926). Experimentelle Untersuchung über phänomenale Identität. *Psychologische Forschung, 7*, 81–136.

Tsuchiya, N., & Koch, C. (2004). Continuous flash suppression. *Journal of Vision*, *4*(8), 61-61.

Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., & Schwarzbach, J. (2003). Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, *100*(10), 6275-6280.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366-383.

Wheatstone, C. (1838). Contributions to the physiology of vision.--Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, 371-394.

Wolfe, J. M. (1984). Reversing ocular dominance and suppression in a single flash. *Vision Research*, *24*(5), 471-478.

World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Jama*, *310*(20), 2191-2194.

Wutz, A., Drewes, J., & Melcher, D. (2016). Nonretinotopic perception of orientation: Temporal integration of basic features operates in object-based coordinates. *Journal of vision*, *16*(10), 3-3.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron*, *47*(1), 143-153.

Zhou, H., Friedman, H. S., & Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, *20*(17), 6594-6611.

Chapter 3 Visual crowding

In crowding, target perception strongly deteriorates when presented within the clutter. Crowding is a major challenge for vision because elements are never presented alone. Therefore, crowding is a perfect testbed for investigating spatial integration. The classic model of vision (pooling), which integrates visual features in a hierarchical and feedforward process, was suggested to explain the visual crowding process. However, it could not explain uncrowding, in which the additional flankers improved the performance, because of its feedforward and hierarchical nature. On the other hand, feedforward convolutional neural networks (ffCNNs) are suggested to account for the complex global shape computation. In the first study, I showed that ffCNNs could not reproduce human performance because of their feedforward nature.

Then, I analyzed the configurations of (un)crowding to see what matters. In the second study, I examined the classic crowding situation, where flankers within the crowding window crowd the target. I asked how the target-related information integrates with flankers and showed that the target-flanker integration is not linear but a complex process. Then, in the third study, I dissected the (un)crowding configurations to investigate to what extent low-level features, such as line-line interactions and orientations, matter. I showed that low-level features have a minor impact, but rather the holistic aspects, such as the Gestalt principle of Prägnanz, matter. Finally, in the last study, I created configurations following ten basic Gestalt principles and showed the Gestalt principles could not fully explain (un)crowding. Instead, I showed that subjective grouping and segmentation measures matter.

3.1 Feedforward Convolutional neural networks cannot explain (un)crowding

Full citation: 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.

Summary:

A simple pooling process cannot explain (un)crowding. It is because by the pooling process, adding flankers can only deteriorate performance, stronger crowding. Feedforward Convolutional Neural Networks (ffCNNs) became state-of-the-art models both in computer vision and neuroscience. It is suggested that ffCNNs are comparable to human visual processing, and ffCNNs may account for human-like global shape computation despite their hierarchical and feedforward nature. Here, we claim that ffCNNs' non-linearities and inductive biases trained with millions of images cannot explain (un)crowding, because of their feedforward nature. Since the target information is irretrievably lost in the early pooling process, higher-level neurons have no chance to access the target information.

We used three different pre-trained ffCNNs (AlexNet, conventional ffCNNs trained on ImageNet; ResNet50, a more sophisticated architecture trained on ImageNet; and Geirhos et al. (2018)'s shape-biased ResNet50, and the ResNet50 architecture trained on a dataset tailored to bias the network towards global shape computations) and tested whether ffCNNs can reproduce human-like performance. Unlike humans, all ffCNNs showed crowding but *not un*crowding. Moreover, we observed that only elements in a local region around the target matter for classification. Occluding the vernier target deteriorates the performance, and occluding parts of the flanker surrounding the vernier improves performance. Occluding other parts of the stimulus, however, does not generally affect performance. The same results held for all three ffCNNs.

These results suggest the following. First, ffCNNs trained with large-scale image sets cannot carry out human-like (un)crowding performance. Second, these qualitatively similar results in three different ffCNNs show that using a more sophisticated ffCNN (i.e., ResNet50) does not allow ffCNNs to explain global uncrowding effects. Finally, 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*.

Postprint of the article published in *Vision research*:

Crowding Reveals Fundamental Differences in Local vs. Global Processing in Humans and Machines

Adrien Doerig[†], Alban Bornet[†], Oh-Hyeon Choung, Michael H. Herzog

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 (Krizhevsky, Sutskever, & Hinton, 2012) and segmentation (Girshick, Radosavovic, Gkioxari, Dollár, & He, 2018), to image synthesis (Goodfellow et al., 2014; Karras, Laine, & Aila, 2018) and scene understanding (Eslami et al., 2018). ffCNNs and the human visual system share several similarities. For example, after training on complex visual datasets such as ImageNet (Deng et al., 2009), ffCNN neural activities show high correlations with human and non-human primate neural activities (Khaligh-Razavi & Kriegeskorte, 2014; Nayebi et al., 2018; Yamins et al., 2014) 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 (Lindsey, Ocko, Ganguli, & Deny, 2019; Zeiler & Fergus, 2014). Because of these similarities, ffCNNs trained on complex visual tasks were proposed as models of the human visual system (Khaligh-Razavi & Kriegeskorte, 2014; Kietzmann, McClure, & Kriegeskorte, 2018; Nayebi et al., 2018; VanRullen, 2017; Yamins et al., 2014). 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 (Baker, Lu, Erlikhman, & Kellman, 2018; Brendel & Bethge, 2019; Doerig, Bornet, et al., 2019; Kim, Bair, & Pasupathy, 2019).

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. (2018) created a new dataset in which textural information was of no avail for object recognition. They used a textural algorithm (Gatys, Ecker, & Bethge, 2016) 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; He, Zhang, Ren, & Sun, 2016) 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; reviews: Herzog, Sayim, Chicherov, & Manassi, 2015; Levi, 2011; Whitney & Levi, 2011). This phenomenon is strongest in the periphery, but also occurs in the fovea (Malania, Herzog, & Westheimer, 2007; Sayim, Westheimer, & Herzog, 2010) . 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; Manassi, Lonchampt, Clarke, & Herzog, 2016; Manassi, Sayim, & Herzog, 2012). 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 (Herzog & Fahle, 2002; Pachai, Doerig, & Herzog, 2016; Sayim et al., 2010) and is not restricted to vision (Oberfeld & Stahn, 2012; Overvliet & Sayim, 2016). 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 (Doerig, Bornet, et al., 2019; Lonnqvist, Clarke, & Chakravarthi, 2019; Volokitin, Roig, & Poggio, 2017), 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 (Krizhevsky et al., 2012), an ffCNN that was used as a model of the human visual system (Khaligh-Razavi & Kriegeskorte, 2014; Zeiler & Fergus, 2014), ResNet50 (He et al., 2016), a more sophisticated ffCNN, and the shape-biased network by Geirhos et al. (2018). We provide experimental evidence suggesting that it is the *architecture* of ffCNNs that prevent them from performing human-like global computations, and not the training procedure.

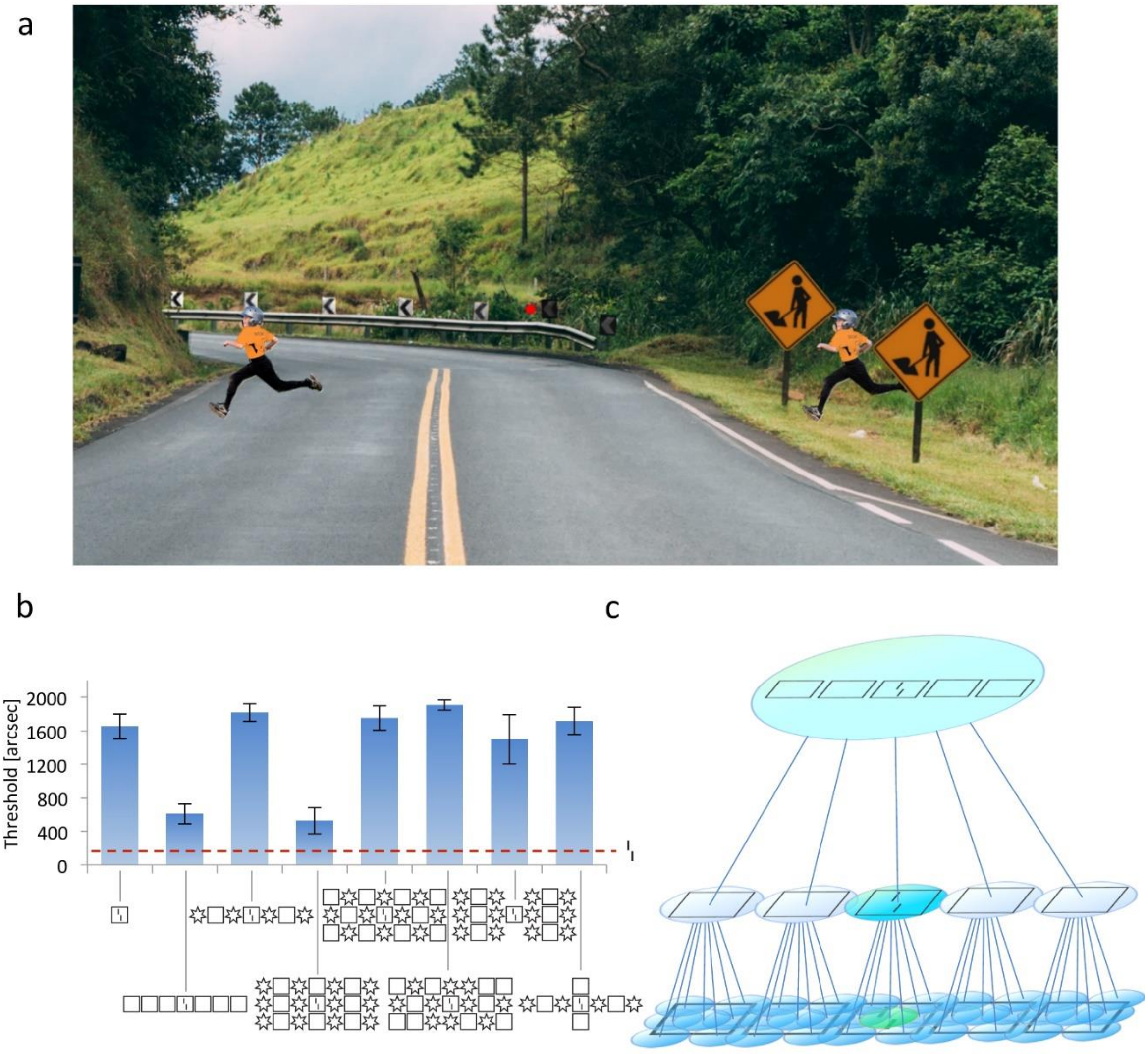


Figure 1. a. In crowding, perception of a target deteriorates in the presence of nearby visual elements. Crowding is ubiquitous in everyday vision, since elements rarely appear in isolation. When fixating on the central red dot, it is more difficult to spot the kid on the right than on the left, because of the nearby signposts. Figure reproduced from Doerig, Bornet, et al. (2019) b. (Un)crowding: Visual elements can be rescued from crowding depending on the global configuration of flankers (*un*crowding). 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, *un*crowding). 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. Modified from Doerig, Bornet, et al. (2019). b. Crowding in ffCNNs: In the feedforward framework of vision, embodied by ffCNNs, crowding occurs by pooling of visual features across a hiererachy 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 sourrouding 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.

Methods

Code and supplementary material are available online at https://github.com/adriendoerig/Doerig-Bornet-Choung-Herzog-2019.

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 (Ioffe & Szegedy, 2015), followed by a hidden layer of 512 units, followed by an ELU non-linearity (Clevert, Unterthiner, & Hochreiter, 2015), 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 (Kingma & Ba, 2014) 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.

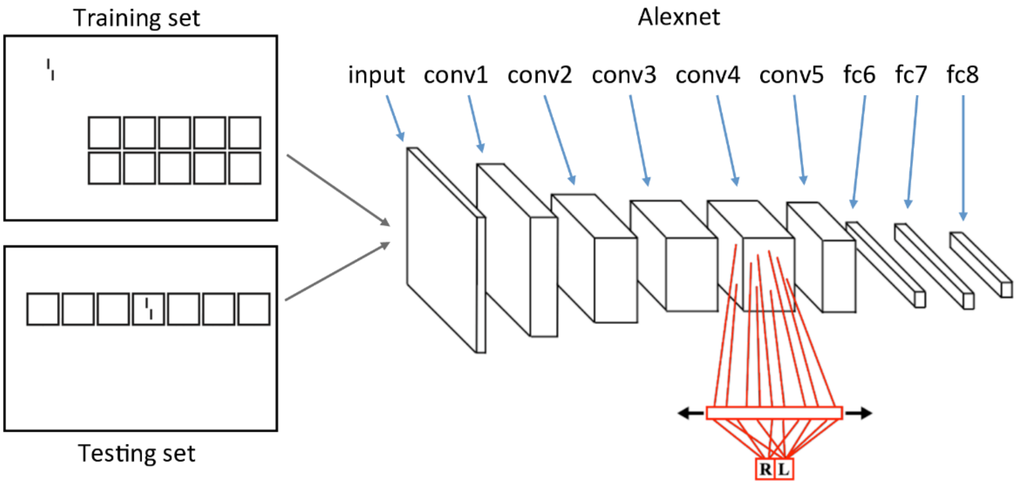


Figure 2. Different stimuli were fed to AlexNet. AlexNet's weights were trained on ImageNet prior to the experiment and were frozen during the experiment. To investigate how well information about the vernier offset is preserved throughout the network hierarchy, we trained one decoder (in red) at each 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 simulatneously 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 Zeiler & Fergus, 2014). 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:

$$score_{P,L} = \frac{\{\vec{T} \cdot (\overrightarrow{y_{P,L}} - \overrightarrow{x_L})\}_{left\_vernier}}{2} + \frac{\{\vec{T} \cdot (\overrightarrow{y_{P,L}} - \overrightarrow{x_L})\}_{right\_vernier}}{2}$$

Where $\overrightarrow{x_L} = (x_1, x_2)_L$ 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), $\overrightarrow{y_{P,L}} = (y_1, y_2)_{P,L}$ 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. (2016): 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 un*crowding. 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 because squares interfere more with verniers due to the vertical orientation of their 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 ffCNNs 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*.

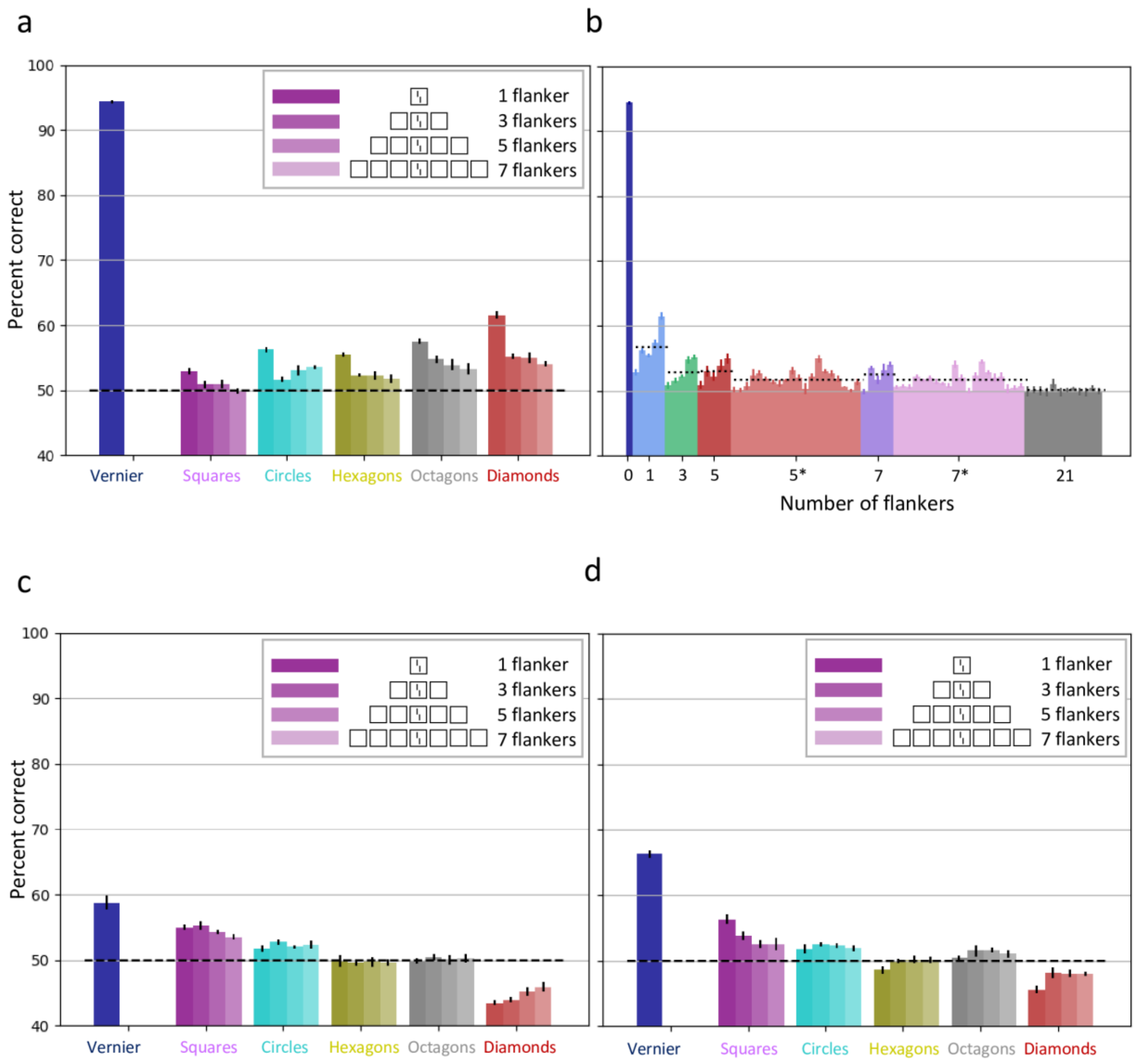


Figure 3. a. Vernier offset discrimination performance for AlexNet with an increasing number of identical flankers. The x-axis shows different flanker configurations. Each color corresponds to one flanker shape, and brighter colors indicate more flankers (from darkest to lightest: 1, 3, 5 & 7 identical 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; Manassi et al., 2016), 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 for (shape-biased) ResNet50 with an increasing number of identical flankers. c. original version. d. Geirhos et al.’s shape-biased version. The results for both of these networks are qualitatively similar for the AlexNet results 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., 2016). 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 (Doerig, Bornet, et al., 2019; Manassi et al., 2016, 2012). 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.

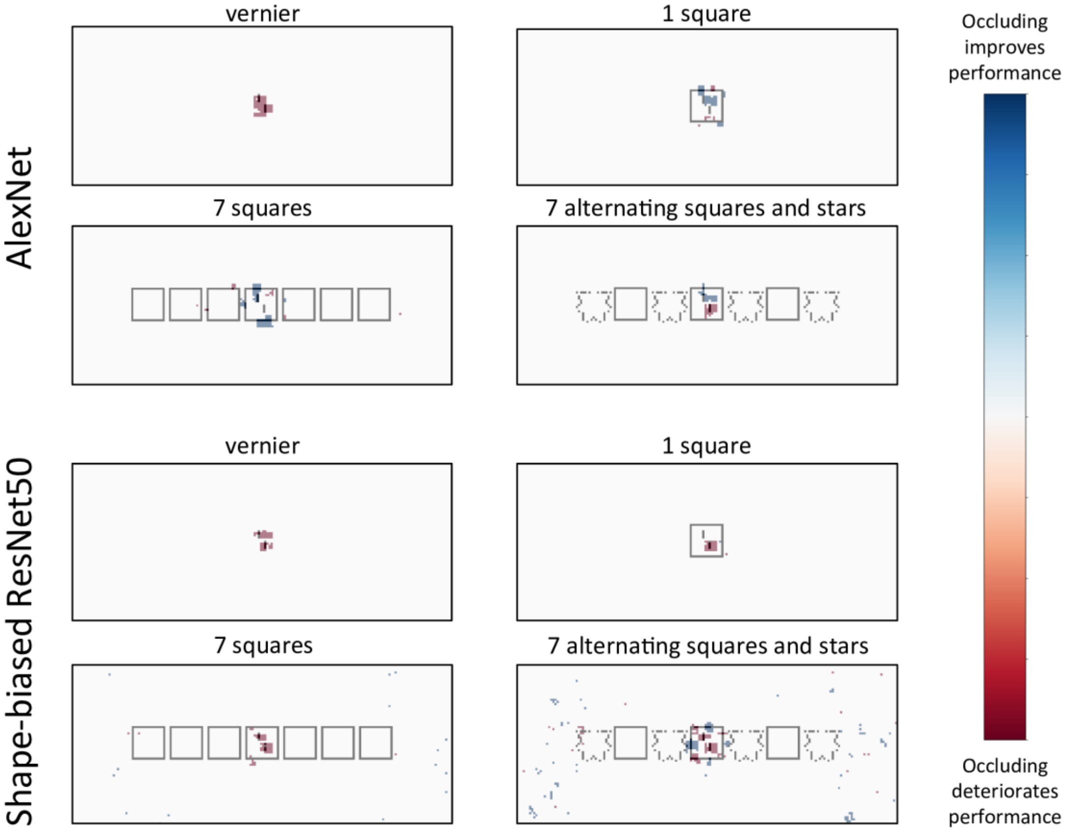


Figure 4. Occlusion analysis. Results of the occlusion analysis for AlexNet (*top*) and the shape-biased ResNet50 (*bottom*). Stimuli on the left lead to good performance in humans, while stimuli on the right lead to strong crowding in humans (Manassi et al., 2016). For both AlexNet and the shape-biased 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 (Manassi et al., 2016; Oberfeld & Stahn, 2012; Overvliet & Sayim, 2016; Whitney & Levi, 2011). 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 (Doerig, Schmittwilken, Sayim, Manassi, & Herzog, 2019).

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 (Doerig, Bornet, et al., 2019; Herzog & Manassi, 2015; Manassi et al., 2016; Pachai et al., 2016). 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 (Doerig, Bornet, et al., 2019; Doerig, Schmittwilken, et al., 2019). 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 (2017) 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., 2019). 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, Schmittwilken, et al. (2019) showed that capsules networks (Sabour, Frosst, & Hinton, 2017), combining CNNs with a recurrent grouping and segmentation process, can explain (un)crowding, including temporal characteristics of uncrowding. Linsley et al. (2018) 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 (Lotter, Kreiman, & Cox, 2016; Nayebi et al., 2018; Spoerer, Kietzmann, & Kriegeskorte, 2019; Spoerer, McClure, & Kriegeskorte, 2017). 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 (Baker et al., 2018; Brendel & Bethge, 2019; Doerig, Bornet, et al., 2019; Doerig, Schmittwilken, et al., 2019; Funke et al., 2018; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Kietzmann et al., 2019; Kim, Linsley, Thakkar, & Serre, 2019; Lamme & Roelfsema, 2000; Linsley et al., 2018; Sabour et al., 2017; Spoerer et al., 2019, 2017; Tang et al., 2018; Wallis et al., 2019). 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. (2018) showed that ffCNNs but not humans are affected by local changes to edges and textures of objects. Brendel and Bethge (2019) 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 (RichardWebster, Anthony, & Scheirer, 2018). 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.

Acknowledgements

Adrien Doerig and Oh-Hyeon Choung were supported by the Swiss National Science Foundation grant n.176153 “Basics of visual processing: from elements to figures”. Alban Bornet was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613.

Bornet, A., Kaiser, J., Kroner, A., Falotico, E., Ambrosano, A., Cantero, K., … Francis, G. (2019). Running large-scale simulations on the Neurorobotics Platform to understand vision-the case of visual crowding. *Frontiers in Neurorobotics*, *13*, 33.

Brendel, W., & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *ArXiv Preprint ArXiv:1904.00760*.

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). *ArXiv Preprint ArXiv:1511.07289*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

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. https://doi.org/10.1371/journal.pcbi.1006580

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2019). Capsule Networks as Recurrent Models of Grouping and Segmentation. *BioRxiv*, 747394. https://doi.org/10.1101/747394

Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., … Gregor, K. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210.

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

Funke, C. M., Borowski, J., Wallis, T. S. A., Brendel, W., Ecker, A. S., & Bethge, M. (2018). Comparing the ability of humans and DNNs to recognise closed contours in cluttered images. *18th Annual Meeting of the Vision Sciences Society (VSS 2018)*, 213.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Preprint ArXiv:1811.12231*.

Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., & He, K. (2018). *Detectron*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672–2680.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Herzog, M. H., & Fahle, M. (2002). Effects of grouping in contextual modulation. *Nature*, *415*(6870), 433. https://doi.org/10.1038/415433a

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.

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.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv Preprint ArXiv:1502.03167*.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974.

Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *ArXiv Preprint ArXiv:1812.04948*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*, 133504.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence required to capture the dynamic computations of the human ventral visual stream. *ArXiv Preprint ArXiv:1903.05946*.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. *ArXiv Preprint ArXiv:1906.01558*.

Kim, T., Bair, W., & Pasupathy, A. (2019). Neural coding for shape and texture in macaque area V4. *Journal of Neuroscience*, *39*(24), 4760–4774.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

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

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Levi, D. M. (2011). Visual crowding. *Current Biology*, *21*(18), R678–R679.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *ArXiv Preprint ArXiv:1901.00945*.

Linsley, D., Kim, J., & Serre, T. (2018). Sample-efficient image segmentation through recurrence. *ArXiv:1811.11356 [Cs]*. Retrieved from http://arxiv.org/abs/1811.11356

Lonnqvist, B., Clarke, A. D., & Chakravarthi, R. (2019). Object Recognition in Deep Convolutional Neural Networks is Fundamentally Different to That in Humans. *ArXiv Preprint ArXiv:1903.00258*.

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *ArXiv Preprint ArXiv:1605.08104*.

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

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. https://doi.org/10.1167/12.10.13

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., … Yamins, D. L. (2018). Task-Driven Convolutional Recurrent Models of the Visual System. *ArXiv Preprint ArXiv:1807.00053*.

Oberfeld, D., & Stahn, P. (2012). Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PloS One*, *7*(10), e48054.

Overvliet, K. E., & Sayim, B. (2016). Perceptual grouping determines haptic contextual modulation. *Vision Research*, *126*(Supplement C), 52–58. https://doi.org/10.1016/j.visres.2015.04.016

Pachai, M. V., Doerig, A. C., & Herzog, M. H. (2016). How best to unify crowding? *Current Biology*, *26*(9), R352–R353. https://doi.org/10.1016/j.cub.2016.03.003

RichardWebster, B., Anthony, S., & Scheirer, W. (2018). Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 3856–3866.

Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychological Science*, *21*(5), 641–644.

Spoerer, C. J., Kietzmann, T. C., & Kriegeskorte, N. (2019). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. *BioRxiv*, 677237.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., … Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840.

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142.

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do deep neural networks suffer from crowding? *Advances in Neural Information Processing Systems*, 5628–5638.

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

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. https://doi.org/10.1016/j.tics.2011.02.005

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833. Springer.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

3.2 How does target information integrate in crowding?

Full citation: Ruthemann, N.*, **Choung O. H.***, & Herzog, M. H., (*in prep*). What crowds in crowding?

Summary:

Crowding has been studied for decades. However, the mechanisms underlying crowding are still largely unknown and controversially discussed. Classically, crowding was explained by local models, where only *neighboring* elements with *similar* features interact with each other, for example, via lateral inhibition. Alternatively, the outputs of the neurons may be pooled, and features may be substituted. Here, we suggest that neither of these models can explain crowding.

We investigated how the target-related information elements within the configuration crowd using the Vernier discrimination task. To do this, we alternated some straight-line flankers in the baseline condition to target-related information; the Verniers have the same (pro-vernier; PV) or opposite (anti-vernier; AV) offset compared to the target Vernier. The baseline condition was composed of one Vernier target in the middle and seven straight lines on each side of the target. We varied the positions of PVs and AVs or the number of PVs and AVs. We expected the PV or AV to impact more when PVs or AVs are placed closer to the target and when more PVs or AVs are presented. However, this was not the case. First, the positions of PVs or AVs did not impact the performance significantly. Interestingly, the number of PVs or AVs affected the performance only in certain conditions. In the other conditions, the number of PVs or AVs did not improve nor deteriorate the performance.

These results were not expected, and we do not have a clear explanation. However, our results suggest that crowding is not a simple linear integration process, but rather a complex process whereindividual strategies play a crucial role.

Manuscript in preparation:

What crowds in crowding?

Nadia Ruethemann[1*], Oh-hyeon Choung[1*], Michael H. Herzog[1]

[1.] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, http://lpsy.epfl.ch, nadia.ruthemann@epfl.ch, ohhyeon.choung@gmail.com

* Authors contributed equally

Abstract :

In visual crowding, perception of an element deteriorates in the presence of clutter. The mechanism(s) underlying crowding are controversially discussed for decades. Whereas it is well established that grouping determines which elements are prone to interference, it is unclear what causes the deterioration. Feature pooling and substitution are competing, but not mutually exclusive, hypotheses. Here, we tested the two hypotheses. We presented an array of lines of which the central target element was offset (vernier offset) and observers were asked to report this offset. The flanking lines were straight or had the same (pro-vernier) or opposite (anti-vernier) offset compared to the target vernier. Participants' performance was hardly influenced by a single pro-vernier but showed a significant decline when a single anti-vernier was presented, contrary to what is expected both from pooling and substitution. However, when the number of pro- as well as anti-verniers increased, performance significantly increased or decreased, respectively. Not only the number of pro-verniers and anti-verniers mattered but also their position. When the flankers were presented closer to the fovea, the influence of the flankers was larger than when presented further away, contrary to the well known in-out anisotropy. In In conclusion, neither pooling nor substitution can explain our results.

Introduction

Perception of a target strongly deteriorates when the target is flanked by neighboring elements. Crowding is a common situation since elements are rarely encountered alone. Crowded features appear as jumbled and their spatial relations are confounded (Levi, 2008). Importantly, features are not rendered invisible as in backward masking for example. Whereas the phenomenology and characteristic of crowding are clear cut, the mechanisms of crowding are discussed for a century (e.g., Adler & Moses, 1965; Ehlers, 1936, 1953; Korte, 1923; Strasburger, 2020; Stuart & Burian, 1962).

Feature pooling is one of the popular explanations of crowding inference, well in the spirit of the classic feedforward model of vision. A stimulus is presented along the visual hierarchy with increasing receptive fields (Hubel & Wiesel, 1962). For example, in V1, the orientation of lines is carefully computed by simple and complex cells having small receptive fields. To detect objects, the outputs of these cells are pooled by neurons in higher level areas, thus almost by definition, having larger receptive fields. This pooling leads to a loss of information evident by jumbled features etc., which deteriorates performance when observers need to report fine grained spatial information. There are two types of pooling. First, in compulsory averaging the reported feature of the target is an average of the features of the target and the flankers (e.g., positions: Dakin et al., 2010; Greenwood et al., 2009; letters: Freeman et al., 2012; orientations: Parkes et al., 2001). By this, the features of the individual target and flankers are irretrievably lost. For example, Parkes et al. (2001) showed that the orientation of a central Gabor target changes depending on the orientation of the surrounding Gabors. The second mechanism is more "pictorial". Features are not averaged but are perceived superimposed and for this reason, performance deteriorates when a response to the target features is asked for. For example, the texture tiling model computes first summary statistics of each feature and returns so-called 'mongrels', which are images where features can be strongly jumbled (Balas et al., 2009; Freeman & Simoncelli, 2011; Rosenholtz et al., 2019). This explanation is well in accordance with the phenomenology of crowding.

A second type of explanation centers around feature specific neural interactions within cortical areas rather than pooling from an area to the next. For example, Sayim and Taylor (2019) found that when several 'T's are presented next to each other, participants report less than the presented number of 'T's. Importantly, the omission rate was worse when the target was highly crowded (Sayim & Wagemans, 2017). Such crowding had only been found when target and flankers were identical and presented repetitively, preferably with a regular spacing (Yildirim et al., 2020). However, such interactions cannot explain averaging.

A third avenue of explanations proposes that features of the target are not averaged with flanker features or jumbled but that the target features are simply substituted by the flanker features for example because of attentional failures (Huckauf & Heller, 2002; Chung, 2002; Strasburger, 2005; Strasburger et al., 1991, Ester et al., 2015). Under substitution, responses follow a bimodal distribution, one peaking around the target feature and the other peaking around the distractor feature (Ester et al., 2014; Gheri & Baldassi, 2008). However, Freeman and colleagues (2012) tested substitution using three alphabet letters: target letter (e.g., '**N**') in the middle, one flanker similar to the target (e.g., '**M**'), and the other very different from the target (e.g., '**I**'). The authors supposed that response errors should be equally distributed on both, flankers similar and dissimilar to the target. Nonetheless, the response errors were largely biased towards target-similar flankers, which is in favor of the pooling mechanism (but see Bernard & Chung, 2011; Coates et al., 2019).

A fourth explanation is that the bottleneck in crowding is not in the feedforward processing stream. All information is processed carefully, but the read out is compromised by bottlenecks in corresponding visual areas, such as in face, motion, emotion, object, related areas, etc. (Farzin et al., 2009; S. He et al., 1996; Louie et al., 2007; Manassi & Whitney, 2018; Wallace & Tjan, 2011).

All these explanations are not mutually exclusive. They might operate in parallel and there are indeed models that tried to come up with unified explanations (Bergen & Landy, 1991; Harrison & Bex, 2015; but see Pachai et al., 2016).

Most explanations have in common that they are restricted to low level feature interactions and locally confined in accordance with Bouma's law (Bouma, 1973; Bouma & Andriessen, 1968), which states the feature interactions occur only within a window of half eccentricity of target presentation.

It is now well established that all these feature-specific explanations have limited explanatory power because they focus on low-level interactions, which is not in line with previous (Chicherov et al., 2014; Chicherov & Herzog, 2015; Doerig et al., 2019; Herzog et al., 2015, 2016; Herzog & Manassi, 2015; Malania et al., 2007; Manassi et al., 2012, 2013, 2015, 2016; Saarela et al., 2009; Sayim et al., 2008, 2010, 2011) and rather recent results (Choung et al., 2019, 2021; Doerig et al., 2019). Here is a simple example. A vernier target is presented and performance is good. Performance deteriorates when the target Vernier is flanked by two neighboring lines (figure 1). This result can be explained by all explanations above. However, when additional lines are added, making the flankers a rectangle, performance strongly improves, leading to uncrowding - even though the deleterious flanking lines are still at the very same position having the very same features. Obviously, there are no low-level interactions in place. None of the above explanations can explain this result (Sayim et al., 2010). Likewise, feature similarity is important but is of no avail when configurations become more complex. What matters in crowding is the overall configuration (e.g., Livne & Sagi, 2007; Manassi et al., 2016; Põder, 2007; Saarela et al., 2009; Sayim et al., 2010; Yeotikar et al., 2011).

We proposed (Herzog et al., 2016; Herzog & Manassi, 2015)that crowding occurs only when elements belong to one group, which may change when additional elements are added, leading to more (Vickery et al., 2009) or less crowding (Manassi et al., 2012, 2013, 2015, 2016 ; Choung et al., 2019, 2021). However, the grouping account does not tell why there is interference at all. Obviously, elements may group, but there is no crowding if the elements are well separated in space.

Here, we used a Vernier offset discrimination task with simple line flankers, which is particularly suited to test for pooling and substitution, to test crowding mechanisms within the group.

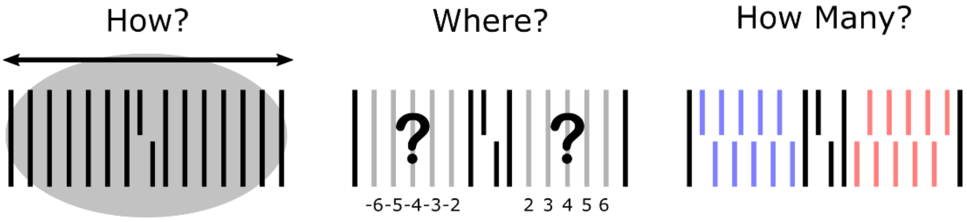


Figure 1. Objective. How the elements within the group interacts. Does the position of congruent or incongruent flanker matter? Does the number of flankers matter? In order to investigate how the elements within the configuration (group) crowd, we compared the baseline condition with the other conditions. Baseline condition was composed of one vernier target in the middle and 7 straight lines on each side of the target. The other conditions included target congruent (pro-vernier, illustrated as blue) and/or target incongruent (anti-vernier, illustrated as red), and we varied the position and the number of pro- and/or anti- verniers.

Materials and Methods

Participants

52 participants took part in four experiments (Exp1: 18 (2 participants excluded from 20 initially recruited participants), Exp2: 11, Exp3: 11, Exp4: 10). Two participants were excluded because they did not show strong crowding in the basic line flanker condition, which is a prerequisite to test for crowding. Hence, we retained the data of 50 participants (mean age: 19.78±1.81, 12 females, 4 left-handed, 23 with left eye dominance). All participants had normal or corrected to normal visual acuity in the Freiburg Visual Acuity Test as indicated by a binocular score greater than 1.0 (Bach, 1996). Observers gave written consent before the experiment. All experiments were conducted in accordance with the Declaration of Helsinki ("World Medical Association Declaration of Helsinki," 2013) and were approved by the local ethics committee (Commission d'éthique du canton de Vaud).

Apparatus

Stimuli were displayed on a gamma-calibrated 24 inch ASUS VG248QE LCD monitor (1920x1080 px, 120 Hz). The room was dimly illuminated (0.5 lux). Viewing distance was 75cm and the participant's chin and forehead were positioned on a chinrest. Responses were collected using hand-held push buttons.

Stimuli

In Experiments 1-3, stimuli were white (100 cd/m^2), presented on a black background with luminance below 0.3 cd/m^2. In Experiment 4, stimuli consisted of (physically) isoluminant red and green lines. In this experiment, the luminance of the vernier target and the flankers was set to 10cd/m^2. Participants were asked to fixate on a white fixation dot (diameter of 8 arcmin, 100 cd/m^2). Stimuli were presented for 150ms. When no response was registered within 3 seconds, the trial was dismissed. A feedback tone was given for omissions (300 Hz) but not for incorrect responses. Vernier stimuli were composed of two vertical bars. Each bar was 40 arcmin long, 1.8 arcmin wide (anti-aliased), and vertically separated by a 4 arcmin gap. The left/right offset size of the vertical verniers was calibrated for each participant (see *Calibration*). Flankers were comprised of 14 verniers with the same (pro-vernier) or different (anti-vernier) offset as the target or were straight lines. Here, we annotated the flanker position depending on the target position. For example, the position on the exact left side of the target is denoted as the -1st position and the far-left one as the -7th position. Lines were composed of 84 arcmin long lines, and all flankers were separated by 30 arcmin. Except for the Vernier alone condition, the vernier target was always flanked by 14 flankers (7 on each side of the target). To reduce target-location uncertainty, two lines (pointers; 40 arcmin long each) 133.33 arcmin above and below the center of the target were indicating the position of the target.

In all experiments, each configuration was presented at the center of the screen, and the fixation dot was presented at an eccentricity of 7 degrees to the left in Experiments 1, 3, 4; and at 5 degrees to the left in Experiment 2. The Psychophysics Toolbox was used to present the stimuli (Brainard, 1997; Kleiner et al., 2007; Pelli & Vision, 1997).

Procedures

General procedure. In all experiments, different flanking configurations were tested in blocks of 80 trials. Two conditions which are the vernier target alone (vernier alone) and the target vernier with 14 straight line flankers (baseline) conditions were tested in all experiments as control configurations (Figs. 2, 3, 4, and

5; grey dotted line: vernier alone condition; red dotted line: baseline condition). Participants were asked to report the vernier offset direction, whether the lower bar was offset to the left (left button) or right (right button) compared to the upper bar. In every block of 80 trials, the number of left and right offsets was balanced. The same offset side was not displayed more than four times in a row.

Calibration. We calibrated the vernier offset size for each participant and used the calibrated offset throughout the experiment, i.e., for the different flanking configurations. The threshold of 70% correct responses in baseline (14 straight lines, 7 on each side of the target) condition were determined by the stair case PEST procedure (Taylor & Creelman, 1967). The starting offset was 16.66 arcmin. In order to avoid extremely large offsets, we restricted the adaptive procedure to 33.32 arcmin (i.e., twice the starting value). In total, participants went through two PEST blocks and one block with calibrated offset of 80 trials.

Experiment 1. 20 observers participated. We tested the flanker configurations as shown in Fig. 2. In Condition 1, Vernier flankers either had same (pro-vernier) or different offset (anti-vernier) as the target and were presented in the second position on the left or right side of the target. In Condition 2, 50% of the trials did not have a target vernier but a flanker vernier in the first or second position on the right or left next to the target position.

Experiment 2. In Experiment 2, 11 observers participated. Eight flanker configurations along with 2 control configurations were tested, where instead of line flankers, pro- and anti-verniers (PV, AV) were presented (Fig. 3). Except for the flankers right next to the target and the outermost flankers, all other flankers were composed of PVs and/or AVs. The flankers right next to the target in position 1/-1 and the outermost flankers in position 7/-7 consisted of gap lines (two 40 arcmin long vertical bars with 4 arcmin vertical gap between the two bars) instead of straight lines like in the other three experiments.

Experiment 3. 11 observers participated and were tested. Additionally, to the 2 control configurations, we added two conditions with 10 and 13 configurations, respectively (Fig. 3). In Condition 1, we varied the position of the anti-vernier flanker (2nd to 6th position to the left or to the right side of the target). In Condition 2, the amount of anti-verniers were modulated (2-5 on the right, on the left or on both sides).

Experiment 4. 10 observers participated and were tested. Additionally, to the 2 control configurations, we added two conditions with 15 configurations each (Fig. 4). These conditions consisted of a basic line flanker configuration, an all-AV configuration, and an all-PV configuration. The all-AV/all-PV configurations are characterized by the fact that all flankers - except the ones next to the target and the outermost flankers - were AVs/PVs. Instead of presenting the stimuli in white on a black background, we varied the colors. In one condition the flankers were red and the target was green and in the other condition the flankers were green and the target was red. We varied the number of flankers having same color as the target, as presented in Figure 4. The number of same-colored flankers were 0, 2, 4, 10, and 14, and the other flankers out of 14 were resented in the opposite color. The stimuli were randomized within configurations, e.g., stimuli with differently colored line flanker configuration were tested in one session (or block), and the stimuli were presented randomly.

Data analysis

We fitted a cumulative Gaussian function to the data using likelihood analyses and determined the vernier offset for which 75% correct responses were reached (threshold). Psignifit version 2.5.6 (see http://bootstrap-software.org/psignifit/; Wichmann & Hill, 2001a, 2001b) was used for the fitting.

Using R (R Core Team, 2019) and *lme4* package (Bates et al., 2015), we computed linear mixed-effects models (LMM) to account for random variations due to individual differences. The fixed and random effects are specified for each experiment. The model significance (p-value) was obtained through likelihood ratio tests (χ^2) by comparing nested models. For each fitted model, using *MuMIn* package (Barton, 2020), we computed the effect size (r^2), i.e., the explained variance, when including (conditional r_c^2) and excluding (marginal r_m^2) the random effects (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013).

Results

Experiment 1. Crowding trend under target congruency and positional uncertainty

Under either pooling or substitution perspective, performance changes when target congruent or incongruent flankers are presented. In our stimuli (Fig. 2), we alternated one of the 14 line flankers a vernier flanker which offset to the same direction (PV) or the opposite direction (AV). When a PV is presented, performance should increase because the combination of target and flanker signals increases. With an AV, performance should deteriorate because of the conflicting information. Similarly, when the target is omitted (Fig. 3), both pooling and substitution models predict the same; a performance deterioration depends on the flanker vernier. Accordingly, two types of flanker configurations were tested: line flankers (baseline condition) and a PV/AV at the second position to the right or left side of the target (Fig. 2), no vernier target but a vernier in the first or second position next to the target position (Fig. 3).

Crowding occurred when the Vernier was flanked by straight lines (Fig. 2, gray dotted line vs. red dotted line, $t(17) = 5.87$, $p < 0.05$). In the PV/AV condition, performance was the worst when an AV was presented in the second position on the left side. In the no vernier target condition, performance was almost the same as performance for the basic line flanker condition. We analyzed these two conditions separately.

In Condition 1, we found a behavioral trend showing that PVs do not have an influence on performance compared to the basic line flanker condition. However, when AVs were displayed, performance decreased compared to the basic line flanker condition.

To summarize, there was no significant difference found between the AVs/PVs and the basic line flanker condition. However, we observed a trend of performance deterioration with anti-verniers.

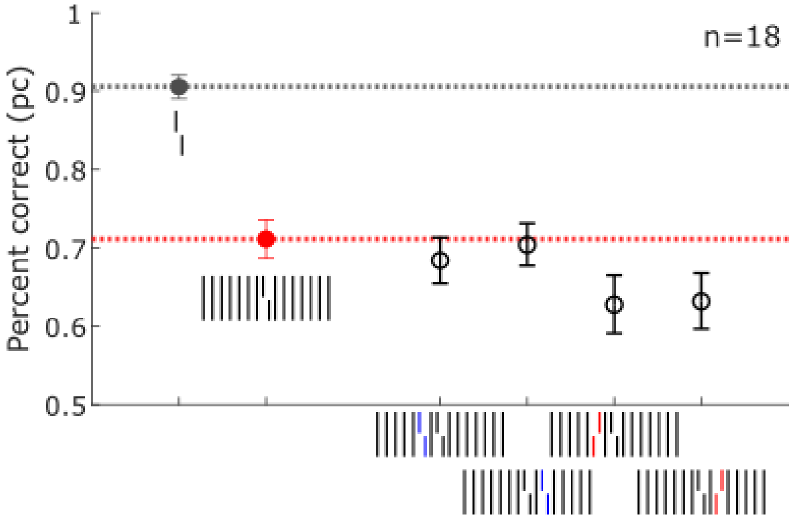


Figure 2. Experiment 1, Condition 1. (±SEM). The gray dotted line represents the Vernier only condition. The red dotted line represents the line flanker condition. The y axis represents Vernier offset discrimination performance in percent correct (±SEM), and the x axis represents the position of the AV (in red) or PV (in blue). Blue represents PVs, red represents AVs. An AV/PV was shown in the second position either on the left or the right side of the target.

To analyze Condition 2, we computed correlations between performance in each configuration and the basic line flanker configuration. The results showed that regardless the flanking vernier position, the performance of the basic line flanker configuration was significantly correlated with that of the no-vernier target configurations (basic line flanker vs. vernier 2nd left; *t*(16)=3.489, $r = 0.657$, $p < 0.05$; basic line flanker vs. vernier 1st left; *t*(16)=2.770, $r = 0.569$, $p < 0.05$; basic line flanker vs. vernier 2nd right; *t*(16)=2.642, $r = 0.551$, $p < 0.05$; basic line flanker vs. vernier 1st right; *t*(16)=4.774, $r = 0.657$, $p < 0.05$). Moreover, the performance of no-vernier target configurations could be predicted by the performance of the basic line flanker configuration by a simple linear regression, and the estimates were close to 1 ($\beta = 0.745 \pm 0.138$). Thus, the high correlation between the basic line flanker condition and the no target vernier conditions and the estimates close to 1 show that humans are able to pick up signals even if the Vernier is not presented in the cued target position.

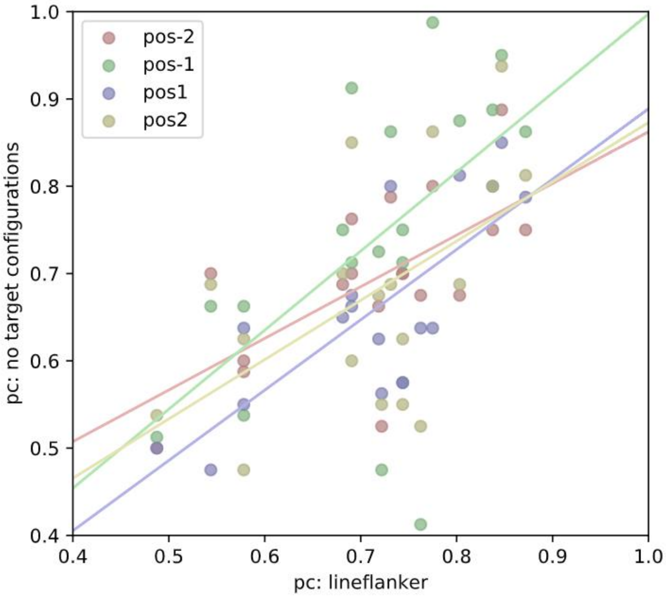


Figure 3. Experiment 1, condition 2. There are strong linear correlations between basic line flanker condition and no-target conditions. X-axis represents performance of basic line flanker condition, and y-axis represent performances of 4 no-target conditions. Each dot represents the individual

performance, and each color represents different configurations; red is the conditions with the flanker Vernier on 2 positions left from the center; green is when the flanker Vernier on 1 position left from the center; purple is when the flanker Vernier on 1 position right from the center; and yellow is when the flanker Vernier on 2 positions right from the center. Each colored line is the linear fitting line between the basic flanker condition and the no-target conditions.

Experiment 2. Combinations of AVs and PVs deteriorate or improve performance

As shown in Experiment 1, only AVs, but PVs, showed a trend to affect the performance. However, in Experiment 1, only one AV/PV was presented among other flankers. In experiment 2, we were therefore interested in the effect of a larger number of AVs and PVs. Thus, 8 configurations with a varying number of AVs and PVs as flankers were tested. Another difference to Experiment 1 was, that the flankers right next to the target were gap lines instead of straight lines. Again, we expected that more PVs improve performance and more AVs deteriorate performance. In this experiment, the results partially agreed with the expectations; PVs improved, and AVs deteriorated the performance, however, only the PVs and AVs on the left (Fig. 3.1.2).

Performance was worse the more AVs are on the left side (Fig. 2). To analyze the relation between performance and number of AVs, we computed an LMM with the number of AVs on the left side and right side as fixed effects. For example, the condition with all of the flankers being AVs was coded as having 5 AVs on the right side and 5 AVs on the left side. Note that the number of AVs on each side also represent the number of PVs on each side, i.e., the number of PVs on the left is equal to 5 - the number of AVs on the left. Individual observers were considered as random intercepts. We found no significant interaction between the two fixed effects (likelihood ratio test between an additive and an interaction model: $\chi^2(1)=0.0013$, p=0.971). Only the fixed effect AV left showed a significant difference, but not the fixed effect AV right (AV left: $\chi^2(1)=53.658$, p<0.05; AV right: $\chi^2(1)=2.093$, p=0.148). The negative parameter estimates in both terms (in Supp. Table 2) show that the percentage of correct responses significantly decreased when the number of AVs on the left side increased, which means performance decreased. The model explains 74.5% of the variance, and only 23.4% when not accounting for the random effects (rm2=0.234, rc2=0.745).

In summary, performance significantly decreases with the number of AVs on the left side but not on the right side. Thus, target-alike flankers presented closer to the fovea have a stronger influence on crowding.

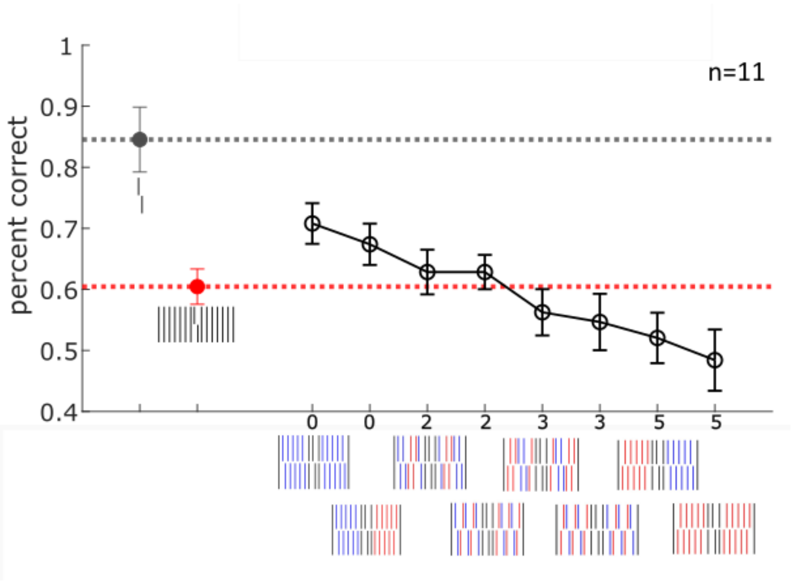


Figure 4. Experiment 2. The y-axis shows the mean percent correct (±SEM). The x-axis shows the configurations. The numbers below the configurations indicate how many AVs were presented on the left side of the target (the significant fixed effect for the LMM). The more AVs were on the left side, the lower was the percent correct.

Experiment 3. Position and number of AV influence performance

In Experiment 1 and 2 we showed that mainly AVs have an influence on target discrimination, hence, we investigated in which spatial and quantitative extent the AVs impact the performance. We tested two conditions. In one condition we varied the position of the AVs and in the other condition we increased the number of AVs either on the left, the right or both sides of the target (see Fig. 5). Both models predict a gradual performance deterioration as AVs are positioned further away from the target. Also, when the number of AVs increases in any direction, the more the AVs are, the worse the performance should be. Results were mixed; performance deteriorated by adding AVs, but we did not observe a clear linear relationships between the positions of AVs or the number of AVs and the performance.

To analyze the above-mentioned relations (Position of AV – percent correct, Number of AV – percent correct), we computed two separate LMMs, one for the position and one for the number. The LMM for the position consisted of the position of the AV as a fixed effect and the individual observers as random effect. The fixed effect showed a significant difference (position of AV: χ^2(1)=6.8152, p<0.05). The positive parameter estimate (in Supp. Table 3) shows that percent correct significantly increased, the more the AV was presented on the right side. Thus, the closer the AV was presented to the fovea, the worse was performance. The model explains 68.2% of the variance, and only 2% when not accounting for the random effects (rm2=0.021, rc2=0.682). However, performance did not deteriorate gradually depending on the AV position. Fig. 5 'Position of AV' columns show that performances are equally bad when an AV was presented on the left side, the parameter estimate was close to zero (β=0.003), inferring that the effect of the AV position was not gradual.

In the LMM for the number, the fixed effects were the number of AVs and the side on which it had more AVs. Individual observers were considered as random intercepts. We found no significant interaction between the two fixed effects (likelihood ratio test between an additive and an interaction model: χ^2(2)=1.0929, p=0.579). Only the fixed effect number of AV showed a significant difference, but not the fixed effect side of AV (num of AV: χ^2(1)=4.1709, p<0.05; side of AV: χ^2(2)=2.7285, p=0.256). The negative parameter estimats in the number of AVs (in Supp. Table 4) shows that percent correct significantly decreased when the number of AVs increased, which means performance decreased. The model explains 63.9% of the variance, and only 1.5% when not accounting for the random effects (rm2=0.015, rc2=0.639). However, there was no gradual performance deterioration, the estimates were close to zero (β=-0.005).

To summarize, the number of AVs as well as the position of AVs has a significant influence on performance. The higher the number of AVs, the worse is performance and the further to the left (closer to the fovea) the AVs are present, the worse is performance.

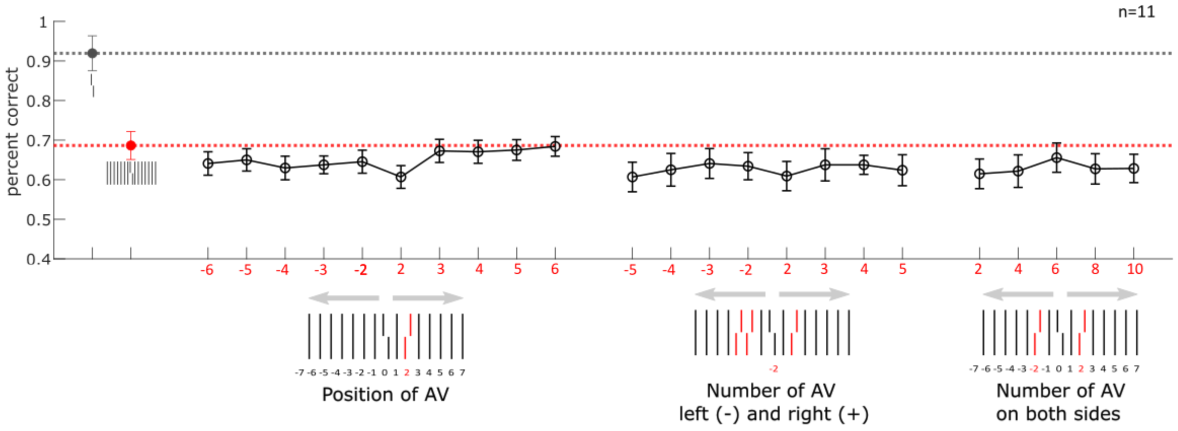


Figure 3. Experiment 3. The y-axis shows the mean percent correct (±SEM). The x-axis shows the configurations. The gray dotted line represents the Vernier alone condition and the red line represents the basic line flanker condition. In the left side, performance in Condition 1 is presented. The numbers indicate the 'position of AV', from -6 (6th position on the left) to +6 (6th position on the right). In the middle and on the right side, Condition 2 is presented, where the number of AVs increases. In the middle, the number of AVs only increases on one side. The numbers incidate how many AVs there are on the left side (indicated with -) and how many there are on the left side (indicated with +). On the right, the numbers indicate how many AVs there are on both sides.

Experiment 4. Grouping effects prevent AVs and PVs from impacting performance

Previous experiments showed that the number AVs has an influence on the percent correct. According to previous research (Manassi et al., 2012, 2015; Sayim et al., 2008, 2010), this effect should diminish when target and flankers are segmented to different groups, because of the differently colored stimuli (see Fig. 6). The pooling model suggests only target similar flankers crowds the target, so in the pooling model, only flankers in the same color as the target deteriorate performance. This results in the same behavioral pattern as the grouping interpretation. However, it does not mean only target-similar flankers crowd. For example, Sayim and colleagues (2008) showed that green flankers could crowd the red target. The prediction for substitution model is unclear.

Our results are in line with previous research. Performance was best when the target was standing out from the flankers (in our case in a different color) (Fig. 6). To analyze the relation between the number of colored stimuli and the percentage of correct responses, we computed three LMMs for each condition (basic line flankers, AVs, PVs) with the number of target-colored stimuli and the color (green vs. red) as fixed effects. For example, when only the target had a different color than the flankers, the number of colored stimuli was coded as 1. Individual observers were considered as random intercepts.

Basic line flanker condition. We found no significant interaction between the two fixed effects (likelihood ratio test between an additive and an interaction model: $\chi^2(1)=0$, p=1). Both fixed effects, the number of target-colored stimuli and color showed significant differences for the basic line flanker condition (Num colored stim: $\chi^2(1)=8.414$, $p<0.05$; Color: $\chi^2(1)=4.440$, p=0.035). The negative parameter estimates in both variables (in Supp. Table 5) show that the performance significantly decreased when the number of colored stimuli increased and when the flankers were green instead of red. The model explains 70.6% of the variance, and only 3.9% when not accounting for the random effects (rm2=0.039, rc2=0.706).

The AV condition. We found no significant interaction between the two fixed effects (likelihood ratio test between an additive and an interaction model: $\chi^2(1)=1.747$, p=0.186). Only the number of target-colored stimuli, but not the color showed a significant in the AV condition (Num colored stim: $\chi^2(1)=8.048$, $p<0.05$; Color: $\chi^2(1)=1.486$, p=0.223). The negative parameter estimates in both variables (in Supp. Table 6) show that percent correct significantly decreased when the number of colored stimuli increased and when the flankers were green instead of red. The model explains 41.5% of the variance, and 5.8 % when not accounting for the random effects (rm2=0.058, rc2=0.415).

The PV condition. We found no significant interaction between the two fixed effects (likelihood ratio test between an additive and an interaction model: $\chi^2(1)=0.142$, p=0.706). Neither the number of colored stimuli, nor the color showed a significant main effect in the PV condition (Num colored stim: $\chi^2(1)=1.628$, p=0.202; Color: $\chi^2(1)=0.00$, p=0.98). The parameter estimates (in Supp. Table 7) for the number of target-colored stimuli were negative, meaning that when the target is more salient, performance is better (but not significantly). The parameter estimates for the color were positive, but again without a significant difference. The model explains 71.7% of the variance, and only 0.4 % when not accounting for the random effects ($r_m^2=0.004$, $r_c^2=0.717$).

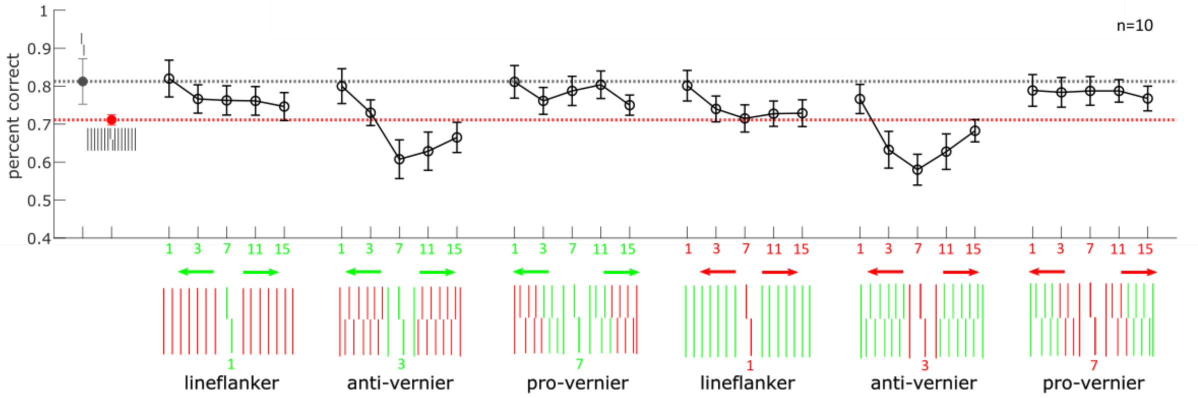


Figure 6. Experiment 4. The y-axis shows the mean percent correct (±SEM). The x-axis shows the different configurations. The numbers in all conditions show how many stimuli were green/red expanding from the center. The stimuli were randomized within configurations, e.g., stimuli with the red target line flanker configuration were tested in one session (or block), and stimuli with different number of red line flankers (0, 2, 6, 10, or 14 red line flankers) were presented randomly within the session.

Discussion

Despite the centuries of studies on crowding, the mechanism remains controversial, e.g., pooling versus substitution. Here, we used the simple stimuli in which the target and flankers are presented within Bouma's window to study whether a single mechanism can explain variable conditions. However, the experimental results were highly variable, where no single mechanism is available to explain. Thus we suggest that the competing mechanisms are not mutually exclusive but complementary or rely on idiosyncratic strategies. Alternatively, crowding may rely on a completely different mechanism.

As expected, when line flankers are presented next to the Vernier, performance drops (baseline performance, classic crowding). However, performances were hardly predictable when target-matched (congruent; PV) flankers or target-unmatched (incongruent; AV) flankers alternate the simple line flankers. If pooling or substitution models were to be true, performances would follow the simple rules: 1) AV deteriorates performance even more than line flankers, where PV improves performance; 2) the more corresponding flankers, the more the performance trend follows that direction, i.e., multiple AVs deteriorate performance more than a single AV.

Here are our observations. First, in all experiments, AVs deteriorated the performance. However, PVs improved performances in some cases, whereas not in the other cases. For example, in Exp. 1 condition 1, a single PV on the left or right side of the target did not improve the performance, while AV on either the left or right side showed a trend of performance deterioration (Fig. 2). On the other hand, in Exp. 2, the number of PVs largely improved the performance (Fig. 4 0AV conditions).

Second, the number of AVs indeed deteriorated the performance but did not have a linear relationship. In Exp. 2, increasing the number of AVs on the left side gradually deteriorated the performance, which can be shown with negative beta estimates (β=-0.039). However, in Exp. 3 conditions 2 and 3, despite the fixed effect of the number of AVs being significant, the estimated slope was close to zero (β=-0.005). The near-zero slope shows that crowding or feature integration is not monotonic nor linear summation. The results again go against either the pooling or substitution mechanism.

Third, the position of AVs did not gradually affect the performance. As all flanker positions were within the Bouma's window, AV affected the target discrimination performance. However, the position did not affect, which is against the previous studies' findings (e.g., Andriessen & Bouma, 1976; Bouma, 1973; Toet & Levi, 1992), showing that the closer the flankers were the stronger the crowding was. Moreover, our results of Exp. 3, showed the well-known in-out anisotropy in the opposite pattern. Crowding is known to have in-out anisotropy, that is, the flankers further away from fovea deteriorated the target performance more than the flankers closer to the fovea (Bouma, 1973; Petrov et al., 2007; Petrov & Meleshkevich, 2011; Whitney & Levi, 2011). Motter and Simoni (Motter & Simoni, 2007) argued that due to the logarithmic compression of V1 cortical space with eccentricity (cortical magnification factor) the outward mask is closer to the target than the inward mask on the cortex, which may explain the inward–outward anisotropy of crowding (also see Pelli, 2008). On the other hand, Petrov and Meleshkevich (Petrov & Meleshkevich, 2011) suggest that inward-outward anisotropy is instead a consequence of unequally distributed attentional locus within the visual field. However, Exp. 3 results showed that when AV was presented closer to the fovea (left side, inward location), rather than when AV was presented on the right side (outward location) of the target. Moreover, in Exp.2, the more PVs and the fewer AVs on the left increased the performance, whereas the PVs and AVs on the right side had no clear effect. Interestingly, when comparing two zero left-AV conditions (Fig. 4 columns 3 vs. 4), although five PVs on the right are alternated by AVs, performance did not deteriorate as much. These observations go against the well-known anisotropic effect of crowding, inward-outward anisotropy. This gives another evidence that traditional model of crowding hardly explain crowding.

Forth, interestingly, participants were able to pick up the target information despite being omitted. Both pooling and substitution models predict performance deterioration. In the pooling model, performance drops because the Vernier is not at the pooling center, which diminishes offset information. In the substitution model, performance drops because the Vernier is off-centered, where the substitution probability decreases. However, surprisingly, performance was almost at the same level as the baseline line flanker condition (linear fit with the baseline condition showed near-one slope estimate).

Fifth, we found the grouping effect as shown in previous studies (Manassi et al., 2012, 2015; Sayim et al., 2010). Effect of AVs is diminished when target and flankers are segmented into different groups. The predictions differ for the pooling and substation model. The pooling model suggests only target similar flankers crowds the target, so in the pooling model, only flankers in the same color as the target deteriorate performance. This results in the same behavioral pattern as the grouping interpretation. However, it does not mean only target-similar flankers crowd. For example, Sayim and colleagues (2008) showed that green flankers could crowd the red target. On the other hand, the substitution model suggests crowding does not depend on the target-flanker similarity. Therefore, performance should be similar to the no-color configurations, that is, PVs improve performance, and AVs deteriorate performance.

In addition, we observed an interesting trend, that is, some of the flankers are seemed to be ignored. As mentioned, Sayim and Taylor (2019) suggested that elements can be masked in the redundant context. For example, when 'T's are repeatedly presented ('TTT'), we perceive less 'T's than the presented number of 'T's. The similar process seems happened in out stimuli. For example, in Exp. 3, ten AVs did not deteriorate performance compared to two AVs. This may be because AV signal is suppressed by lateral inhibition. Hence, in the feature space, only a similar amount of antivernier signal is left, leading to similar performance.

Overall, the results are a clear indication, that a lot of processes are going on in the brain during visual crowding. Furthermore, these results show that pooling or substitution is not mandatory but rather that basic object recognition happens at a post-perceptual stage, where individual strategies play a crucial role.

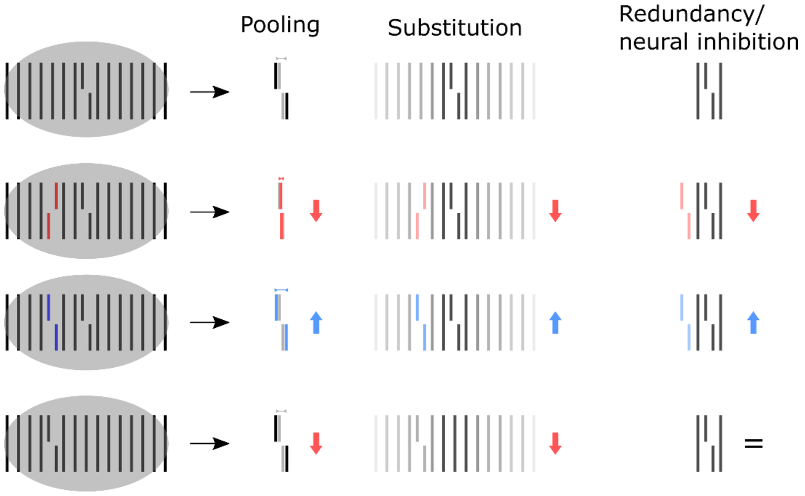


Figure 7. The expected behavior of classic crowding models, when flankers are in the traditional 'crowding window' (Bouma's window, ½ ecentricity). Left most column shows different experimental conditions of experiment 1. 2nd left column shows the predictions from pooling model (averaging pooling). 2nd right column shows the predictions from basic substitution model. Right most column shows the prediction from neural inhibition model (redundancy model). For the visualization purpose, AVs are painted as red, and PVs are in blue. Stimuli were in white color in the experiments. Red downward arrows represent expected performance deterioration compared to the baseline condition (1st row), blue upward arrows represent expected performance improvement, and '=' represents similar expected performance.

Acknowledgement

NR, OHC, MHH were supported by the Swiss National Science Foundation (SNF) 320030_176153 "Basics of visual processing: from elements to figures".

References

Allik, J., Toom, M., Raidvee, A., Averin, K., Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research, 83*, 25-39. https://doi.org/10.1016/j.visres.2013.02.018

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

Awh, E., Matsukura, M., & Serences, J. T. (2003). Top-down control over biased competition during covert visual orienting. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 52-63. https://doi.org/10.1037/0096-1523.29.1.52

Bach, M. (1996). The Freiburg Visual Acuity Test—Automatic Measurement of Visual Acuity. *Optometry and Vision Science*, *73*(1), 49–53.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12):13, 1-18. https://doi.org/10.1167/9.12.13

Banks, W. P., Larson, D. W. & Prinzmetal, W. (1979). Asymmetry of visual interference. *Perception & Psychophysics, 25*, 447-456.

Barton, K. (2020). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1--48. https://doi.org/10.18637/jss.v067.i01

Bouma, Herman. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research*, *13*(4), 767–782.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Dakin, S., Cass, J., Greenwood, J., & Bex, P. (2010). Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli. *Journal of Vision, 10*(10): 14, 1-16. https://doi.org/10.1167/10.10.14

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*, 415-434. https://doi.org/10.1016/j.neuron.2012.01.010

Dosher, B. A., & Lu, Z. L. (2000). Mechanisms of perceptual attention in precuing of location. Vision Research, 40, 1269-1292. https://doi.org/10.1016/S0042-6989(00)00019-5

Ester, E. F., Klee, D., Awh, E. (2014). Visual crowding cannot be wholly explained by feature pooling. *J. Exp. Psychol. Hum. Percept. Perform., 40*(3), 1022-1033. https://doi.org/10.1037/a0035377

Ester, E. F., Zilber, E., Serences, J. T. (2015). Substitution and pooling in visual crowding induced by similar and dissimilar distractors. *Journal of Vision, 15*(1):4, 1-12. https://doi.org/10.1167/15.1.4

Fischer, J., Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, *106*(3), 1389-1398. https://doi.org/10.1152/jn.00904.2010

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Attention, Perception, & Psychophysics*, *74*, 379-396. https://doi.org/10.3758/s13414-011-0229-0

Freeman, J., Pelli, D. G. (2007). An escape from crowding. *Journal of Vision, 7*(2): 22, 1-14. https://doi.org/10.1167/7.2.22

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6), 16–16. https://doi.org/10.1167/11.6.16

Greenwood, J., Bex, P., & Dakin, S. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences, USA, 106*, 12130-13135. https://doi.org/10.1073/pnas.0901352106

Greenwood, J. A., & Parsons, M. J. (2020). Dissociable effects of visual crowding on the perception of color and motion. *Proceedings of the National Academy of Sciences*, *117*(14), 8196–8202.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature, 383*, 334-337. https://doi.org/10.1038/383334a0

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.

Herzog, M. H., Thunell, E., & Ögmen, H. (2016). Putting low-level vision into global context: Why vision cannot be reduced to basic circuits. *Vision Research*, *126*, 9–18. https://doi.org/10.1016/j.visres.2015.09.009

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in cat's visual cortex. *The Journal of Physiology, 160*, 106-154. https://doi.org/10.1113/jphysiol.1962.sp006837

Huckauf, A., & Heller, D. (2002). What various kinds of errors tell us about lateral masking effects. *Visual Cognition*, *9*(7), 889–910.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cognitive Psychology, 43*(3), 171-216. https://doi.org/10.1006/cogp.2001.0755

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Larson, A. M., Loschky, L. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):6, 1-16. https://doi.org/10.1167/9.10.6

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. https://doi.org/10.1016/j.visres.2007.12.009

Levi, D. M., Klein, S. A., & Aitsebaomo, A. P. (1985). Vernier acuity, crowding and cortical magnification. *Vision Research, 25*(7), 963-977. https://doi.org/10.1016/0042-6989(85)90207-X

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision*, *7*(2), 4. https://doi.org/10.1167/7.2.4

Loschky, L. C., McConkie, G. W., Yang, J., Müller, M. E. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition*, 12, 1057-1092. https://doi.org/10.1080/13506280444000652

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, 1-11. https://doi.org/10.1167/7.2.24

Malania, M., Herzog, M. H., Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *Journal of Vision, 7*(2), 1-7. https://doi.org/10.1167/7.2.1

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

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. https://doi.org/10.1167/12.10.13

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision*, *13*(13), 10. https://doi.org/10.1167/13.13.10

Martelli, M., Majaj, N. J., & Pelli, D. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision, 5*(1), 58-70. https://doi.org/10.1167/5.1.6

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (n.d.). *The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded*. 11.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology*, *18*(4), 445–451.

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

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

R Core Team. (2019). *R: A Language and Environment for Statistical Computing* (Vienna, Austria). R Foundation for Statistical Computing. https://www.R-project.org/

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.

Rosenholtz, R., Yu, D., & Keshvari, S. (2019). Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of Vision*, *19*(7), 15–15. https://doi.org/10.1167/19.7.15

Sayim, B., Cavanagh, P. (2013). Grouping and crowding affect target appearance over different spatial scales. *PLOS ONE*, 8(8): e71188. https://doi.org/10.1371/journal.pone.0071188

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in vernier acuity. *Journal of Vision*, *8*(8), 12–12. https://doi.org/10.1167/8.8.12

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, 1-23. https://doi.org/10.1167/7.2.7

Sharikadze, M., Fahle, M., & Herzog, M. H. (2005). Attention and feature integration in the feature inheritance effect. *Vision Research, 45*, 2608-2619. https://doi.org/10.1016/j.visres.2005.03.021

Strasburger, H. (2005). Unfocussed spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision*, *5*(11), 8–8. https://doi.org/10.1167/5.11.8

Strasburger, H., Harvey, L. O., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Perception & Psychophysics*, *49*(6), 495–508.

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient Estimates on Probability Functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787. https://doi.org/10.1121/1.1910407

Zhang, J.-Y., Zhang, G.-L., Lei, L., Yu, C. (2012). Whole report uncovers correctly identified but incorrectly placed target information under visual crowding. *Journal of Vision, 12*(7):5, 1-11. https://doi.org/10.1167/12.7.5

Supplemental Tables *Parameter estimates of Linear Mixed Effects Models (LMMs)*

Table 1. Estimates from the linear mixed-effects model of Exp1 of Condition 1 with the position of AV (no main effect) and position of AV (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.672	0.026	26.137
Position of AV	0.001	0.008	0.136
Position of PV	0.005	0.008	0.658

Table 2. Estimates from the linear mixed-effects model of Exp2 with the number of AVs left (main effect) and AVs right (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.708	0.037	19.352
The num of AVs left	-0.039	0.004	-8.808
The num of AVs right	-0.006	0.004	-1.456

Table 3. Estimates from the linear mixed-effects model of Exp3 Condition 1. with the number of AVs (main effect) and the side of AVs (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.651	0.024	27.302
The position of AV	0.003	0.001	2.656

Table 4. Estimates from the linear mixed-effects model of Exp3 Condition 2. with the number of AVs (main effect) and the side of AVs (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.672	0.034	19.477
The num of AVs	-0.005	0.003	-2.057
The side of AVs (left)	-0.021	0.015	-1.400
The side of AVs (right)	-0.021	0.150	-1.381

Table 5. Estimates from the linear mixed-effects model of Exp4 line flanker condition with the number of AVs (main effect) and the side of AVs (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.800	0.034	22.967

The num of colored stim	-0.004	0.001	-2.970
Color (red)	-0.029	0.134	-2.133

Table 6. Estimates from the linear mixed-effects model of Exp4 AV condition with the number of AVs (main effect) and the side of AVs (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.735	0.037	19.782
The num of colored stim	-0.007	0.002	-2.902
Color (red)	-0.029	0.023	-1.224

Table 7. Estimates from the linear mixed-effects model of Exp4 PV condition with the number of AVs (main effect) and the side of AVs (no main effect) as predictors (no interaction between the two predictors) and individual observers as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.794	0.034	23.701
The num of colored stim	-0.002	0.001	-1.282
Color (red)	0.000	0.012	0.020

3.3 Dissecting (un)crowding

Full citation: **Choung, O. H.,** Bornet, A., Doerig, A., & Herzog, M. H., (2021). Dissecting (un)crowding, *Journal of Vision*, 21(10):10, 1-20.

Summary:

In this study, we systematically dissected the holistic configuration of uncrowding, and observed to what extent local and low-level features impact (un)crowding. First, we tested whether parts of the squares, such as line-line detector inhibition by surrounding suppression, or contour-contour interactions induced illusory contours, can explain uncrowding. As a result, the complete square conditions showed better performance than the partial square conditions in multiple square conditions, independently of the number of flankers or flanker orientations. Thus, not line-line detection inhibitions nor colinear contour-contour interaction (illusory contour) could explain uncrowding, suggesting that rather holistic aspects, such as the Gestalt principle of Prägnanz, matter.

Next, we tested whether "crowding of crowding" can fully explain uncrowding. Manassi and colleagues (2013) suggested that uncrowding happens when the target Vernier standout from the flankers, due to the central flanking square being crowded by the other flanking squares. The authors adjusted the width of the central square and asked participants to report the aspect ratio of the central square (the aspect ratio discrimination task). They showed that the center square's aspect ratio is harder to discriminate when the number of squares increases. Here, we tested the Vernier and the aspect ratio discrimination tasks with various configurations, which led to uncrowding. Unexpectedly, the performances of two tasks did not correlate to each other. This result suggests that "crowding of crowding" is not sufficient to explain uncrowding.

Third, we tested whether low-level features of crowding are sufficient to explain uncrowding. Crowding is known to have anisotropies because of its local nature. For example, flankers in the radial orientation interfere stronger than flankers in the tangential orientation (radial-tangential anisotropy). The radial-tangential anisotropy is explained by elliptic receptive fields in early visual areas or uneven sampling density in the early visual cortex. We aligned the square flankers in the vertical (tangential) and horizontal (radial) orientations and tested the Vernier and the aspect ratio discrimination tasks. As a result, in the aspect ratio discrimination task, which tests crowding of the center square, the central square in the horizontally aligned (radial) squares was more crowded than in the vertically aligned (tangential) squares, a radial-tangential anisotropy. However, the performance in the Vernier discrimination task did not show the anisotropy. This indicates that low-level features are not sufficient to explain uncrowding.

Finally, we tested the dissected configurations with three models, which take global aspects into account but are based on different premises. Capsule networks and the Laminart model are two-stage models, in which elements are first parsed into different groups, and then interference occurs only within the groups. Capsule networks group elements on the basis of object-level routing by agreement, whereas the Laminart model groups elements on the basis of low-level features. The TTM model is a one-stage model that pools many low-level features computed over pooling regions whose size grows with eccentricity. Capsule Network reproduced the general human behavior pattern well, that is, performance improved when adding more squares and deteriorated with the partial squares. The Laminart model only partially reproduced the human behavior, that is, performance improved when adding more squares and deteriorated with line-line detection inhibition-like partial square conditions. However, unlike humans, model performance improved with ‘illusory contour’-like partial squares conditions. However, the TTM (one-stage model) could not reproduce the human behavior, that is, adding more squares deteriorated performance, and performances with the partial squares were better than in the complete squares conditions. Hence, in summary, our results favor the two-stage models over the one-stage model.

Overall, we showed that low-level impacts on uncrowding are minor. With the model simulations, we showed that two-stage grouping and segmentation models are needed, which groups elements on the basis of object-level features.

Postprint of the article published in *Journal of Vision*:

Dissecting (un)crowding

Oh-Hyeon Choung[1], Alban Bornet[1], Adrien Doerig[1,2], Michael H. Herzog[1]

[1.] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, http://lpsy.epfl.ch, oh-hyeon.choung@epfl.ch

[2.] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Abstract

In crowding, perception of a target deteriorates in the presence of nearby flankers. Surprisingly, perception can be rescued from crowding if additional flankers are added (uncrowding). Uncrowding is a major challenge for all classic models of crowding and vision in general, because the global configuration of the entire stimulus is crucial. However, it is unclear which characteristics of the configuration impact (un)crowding. Here, we systematically dissected flanker configurations and showed that (un)crowding cannot be easily explained by the effects of the sub-parts or low-level features of the stimulus configuration. Our modeling results suggest that (un)crowding requires global processing. These results are well in line with previous studies showing the importance of global aspects in crowding.

Introduction

In crowding, perception of a target strongly deteriorates when embedded in context (review: Herzog et al., 2016; Levi, 2008; Pelli & Tillman, 2008; Strasburger, 2020). Crowding is the standard situation in everyday vision because elements are rarely encountered in isolation. Crowding is stronger when the target and the flankers share similar features, such as same contrast polarity (Kooi et al., 1994), color (Kennedy & Whitaker, 2010; Põder, 2007; van den Berg et al., 2007), orientation (Andriessen & Bouma, 1976; Parkes et al., 2001; Wilkinson et al., 1997), motion (Bex & Dakin, 2005; Gheri et al., 2007), spatial frequency (Chung et al., 2001; Põder & Wagemans, 2007), etc. It is often argued that only flankers within a certain spatial window (Bouma's window) around the target deteriorate performance (Bouma, 1970; Bouma, 1973; Levi, 2008; Strasburger et al., 1991; Weymouth, 1958). Crowding has specific characteristics. For example, flankers in the radial orientation interfere stronger than flankers in the tangential orientation (radial-tangential anisotropy; Chung, 2013; Greenwood et al., 2017; Kwon et al., 2014; Malania et al., 2020; Toet & Levi, 1992), which was explained by elliptic receptive fields in early visual areas (Hubel et al., 1978; Silson et al., 2018; Toet & Levi, 1992) or by an uneven sampling density in the early visual cortex (Kwon & Liu, 2019; Motter & Simoni, 2007).

Accordingly, crowding is traditionally explained by local, feature-specific interactions between the neural representations of the target and its direct neighbors. For example, neurons sharing the same orientation may interact with each other through lateral inhibition, feedforward pooling, etc. (e.g., Dakin et al., 2010; Greenwood et al., 2009, 2017; Parkes et al., 2001; Pelli, 2008; Rosenholtz et al., 2012; Solomon et al., 2004). In all these models, target information is irretrievably lost at the early stages of visual processing. Thus, this kind of crowding research is very much in the spirit of an atomistic view of visual processing, where basic, local processing precedes more complex processing.

However, all these explanations break down when the target is presented with complex, instead of simple flanker configurations (e.g., Livne & Sagi, 2007; Manassi et al., 2016; Põder, 2007; Saarela et al., 2009; Sayim et al., 2010; Yeotikar et al., 2011). For example, a vertical Vernier (target) is presented, and participants indicate whether the lower segment is offset either to the left or right compared to the upper one (Fig. 1). Performance is good when the Vernier is presented alone but strongly deteriorates when surrounded by a square, a classic crowding effect. Traditionally, the deterioration may be explained by interactions between the vertical lines of the square and the Vernier. However, adding more squares does not further deteriorate performance. Instead, performance improves as more squares are added, approaching the performance level of the unflanked Vernier condition (uncrowding). Manassi and colleagues (2013) proposed that the Vernier target is released from crowding, because the additional flanking squares suppress the central square surrounding the Vernier. Uncrowding effects go well beyond Bouma's window and depend on the configuration of the entire stimulus, across more or less the entire visual field (Chicherov et al., 2014; Chicherov & Herzog, 2015; Doerig et al., 2019; Herzog et al., 2015, 2016; Herzog & Manassi, 2015; Malania et al., 2007; Manassi et al., 2012, 2013, 2015, 2016; Saarela et al., 2009; Sayim et al., 2008, 2010, 2011).

Obviously, local approaches are of no avail. Contextual information across large parts of the visual field needs to be taken into account. Accordingly, models that go beyond spatially confined processing are needed. On the one hand, two-stage models propose that visual elements are first parsed in different groups, and then crowding occurs only within these groups. For example, grouping may arise from the integration of low-level features (*Laminart model:* Francis et al., 2017) or from the competitions between different object-level representations of visual content (*Capsule networks:* Doerig et al., 2020; Sabour et al., 2017). On the other hand, Rosenholtz and colleagues (2019) suggested an one-stage pooling model, the texture tiling model (*TTM*), which may account for the complex effect of global configurations, despite its local nature. The main claim is that pooling a sufficiently large number of low-level image statistics (High Dimensional (HD) pooling) can preserve sufficient information about complex configurations, which can be used at the decision-making stage.

Doerig and colleagues (Doerig, Bornet, et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken, et al., 2020) showed with extensive comparisons that grouping and segmentation processes are crucial for (un)crowding. In contrast, Rosenholtz and colleagues (2019) proposed that HD pooling can explain uncrowding (but see Bornet et al., same volume). However, in both approaches, it is unclear which aspects of the configuration may impact (un)crowding and to what extent low-level interactions within sub-parts can explain complex global processing. Moreover, it is currently unclear to what extent low-level features and properties, such as the target orientation or the radial-tangential anisotropy, contribute to (un)crowding.

To study whether (un)crowding is truly a global phenomenon or instead, it can be explained by how sub-parts of the stimulus interfere, we systematically dissected the holistic configuration, as depicted in Fig. 1A. We tested whether parts of the squares, such as their vertical lines, can explain uncrowding or if global processing of the good Gestalt of squareness is needed (Fig. 1A, experiment 1). For example, line-line detector inhibition (Fig. 1A upper) by divisive normalization may suppress the center square (Carandini & Heeger, 2012; Coen-Cagli et al., 2015). Alternatively, contour-contour interactions (Fig. 1A lower) may create an illusory contour, which can group the flankers together and segment them out from the target (Clarke, Herzog, et al., 2014; Doerig et al., 2019; Francis et al., 2017).

As mentioned, crowding is stronger with flankers in the radial orientation than in the tangential orientation. Here, we tested whether there is such an anisotropy also in uncrowding. We presented arrays of squares in cardinal (experiment 1; Fig. 1B, horizontal arrows) or oblique (experiment 2; Fig. 1B, 45° arrows) orientation, and aligned the squares either along the radial (Fig. 1B left) or tangential (Fig. 1B right) direction. Also, we varied the target orientations (vertical, horizontal, or ± 45°) to further assess low-level flanker-target interactions.

Then, we tested to what extent crowding by the flanking squares on the central square determines crowding by the central square on the Vernier (experiment 3). Does "crowding of crowding" lead to uncrowding? Finally, we tested which modeling approaches best suit these results by comparing models based on grouping and segmentation versus HD pooling.

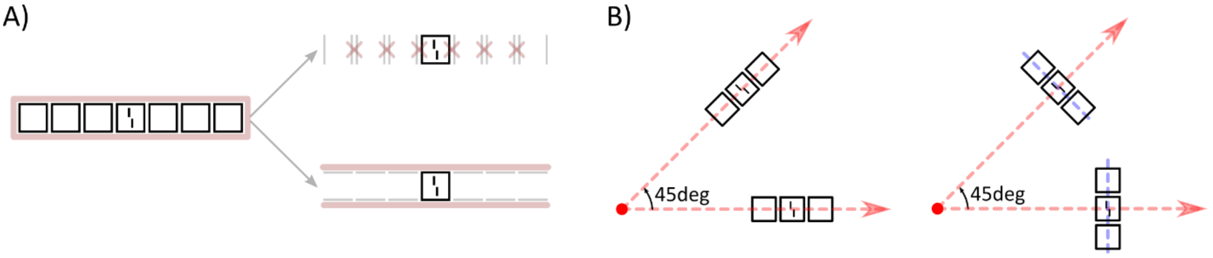


Figure 1. Experimental conditions to test low-level impacts on (un)crowding. A) Experiment 1: Dissecting global configurations to iso-target (upper) and ortho-target (lower) flankers to test if low-level interactions can explain uncrowding. For example, line-line detector inhibitions (iso-target; upper) such as divisive normalization may suppress the center square (Carandini & Heeger, 2012; Coen-Cagli et al., 2015) so that the target uncrowds from the flanker. Alternatively, contour-contour interactions (ortho-target; lower) may create an illusory contour, which can group the flankers together and segment them out from the target (Clarke, Herzog, et al., 2014; Doerig et al., 2019; Francis et al., 2017). B) Experiment 1 & 2: Radial (left) - tangential (right) anisotropic effects on uncrowding either in cardinal (0 °) or oblique (45 °) orientations. Here, red dots represent the fixation point, red dotted line represents the radial axis, and blue dotted line represents the tangential axis.

Materials and Methods

Participants

Thirty-eight participants took part in four experiments (Exp. 1: 11 [one participant was excluded from 12 initially recruited participants], Exp. 2: 10 [five excluded from 15], Exp. 3a: 7 [three excluded from 10], Exp. 3b: 10). Overall, nine participants were excluded right after the calibration session, because they did not show strong crowding in the one square condition, which is a prerequisite to test for a release of crowding to avoid the ceiling effect (see *Calibration session*). Hence, we retained the data of 38 participants from 47 initially recruited participants (mean age: 23 ± 3.7, 17 females, two left-handed, eight with left eye dominance by the Miles test (1930)). All participants had normal or corrected to normal visual acuity in the Freiburg Visual Acuity Test, as indicated by a binocular score greater than 1.0 (Bach, 1996). Participants gave written consent before the experiment. All experiments were conducted following the Declaration of Helsinki except for the preregistration (World Medical Association, 2013) and were approved by the local ethics committee (Commission d'éthique du canton de Vaud).

Apparatus

Stimuli were displayed on a gamma-calibrated 24-inch ASUS VG248QE LCD monitor (1920x1080 px, 120 Hz). The room was dimly illuminated (0.5 lux). The viewing distance was 75cm, and the participant's chin and forehead were positioned on a chin-rest. Responses were collected using hand-held push buttons. In experiment 2, participants' eye movements were tracked with a The Eye Tribe eye tracker (60 Hz sampling frequency, The Eye Tribe, Copenhagen, Denmark), and stimuli were displayed only when participants adequately fixated.

Stimuli

Stimuli were white (100 cd/m^2), presented on a black background with luminance below 0.3 cd/m^2. Participants were asked to fixate on a red fixation dot (diameter of 8 arcmin, 20 cd/m^2). Stimuli were presented for 150ms. When no response was registered within 3 seconds, the trial was repeated randomly within the same block. A feedback tone was given for incorrect responses (600 Hz) and omissions (300 Hz). Vernier stimuli were composed of two vertical/horizontal/45˚ clockwise or counter-clockwise tilted bars (depending on conditions; see below). Each bar was 40 arcmin long, 1.8 arcmin wide (anti-aliased), and separated by a 4 arcmin gap. Left/right offsets of vertical Verniers, up/down offsets of horizontal Verniers, or closer/further from the fixation dot offsets of 45˚ tilted Verniers, were balanced within a block. Flankers were combinations of squares and lines. In the Vernier discrimination tasks in experiments 1, 2, and 3a, the square and the distance between the squares were individually calibrated as described in *Procedures*. Before the calibration, squares and lines were composed of 120 arcmin long lines and were separated by 30 arcmin; thus the center-to-center distance between two flankers was 150 arcmin. For the aspect ratio discrimination tasks in experiments 3a and 3b, stimuli dimensions were identical for all participants; squares and lines were composed of 96 arcmin long lines and were separated by 24 arcmin.

Except for the 45˚ tilted conditions in experiment 2, each configuration was presented at the center of the screen, and the fixation dot was presented at an eccentricity of 9˚ to the left. In the 45˚ conditions of experiment 2, each configuration was presented 2˚ to the right and up from the center. The fixation dot was $\frac{7}{\sqrt{2}}^{\circ}$ to the left and down from the target presentation position. Hence, the target eccentricity was 7˚. Psychophysics Toolbox was used to present the stimuli (Brainard, 1997; Kleiner et al., 2007; Pelli & Vision, 1997).

Procedures

General procedure. Different flanking configurations were tested in blocks of 100 trials. To reduce target-location uncertainty, only the target was presented alone for 150ms at the beginning of each block. We used the PEST (Parameter Estimation by Sequential Testing) stair-case procedure (Taylor & Creelman, 1967). In PEST, test levels are changed step-wise based on the recent response history. The current test level is only changed when the hit rate for this test level lies, with some certainty, above or below the threshold criterion of 75%. The test levels are changed to make the hit rate converge to 75%, thereby boxing the threshold. After a fixed number of trials (100), we ended the procedure and took the threshold from the psychometric function that was fitted to the data post-hoc (details in *Data analysis*). We randomized the order of experimental conditions across participants. In experiments 1, 2, and 3a, participants went through a calibration session to adjust flanker size individually (see *calibration*).

Calibration session. Before the experimental conditions of experiments 1, 2, and 3a (not including Exp. 3b), 37 (Exp. 1: 12, Exp. 2: 15, Exp. 3a: 10) initially recruited participants went through a calibration session to avoid floor and ceiling effects. First, we familiarized participants with the peripheral Vernier task, where only a Vernier target was presented (160 trials, Vernier alone condition). If the Vernier offset threshold was smaller than 200 arcsec, the participant proceeded to the next condition. Otherwise (threshold larger than 200 arcsec), the same block was repeated to familiarize with the stimuli. Thirteen among 37 participants repeated the familiarization block. Second, up to 7 blocks with a Vernier surrounded by one square (80 trials/block) were tested to find the spatial parameters so that thresholds were at least six times larger than in the Vernier alone condition (mean threshold for the Vernier alone condition: 175.9 ± 11.2 arcsec, for the one square condition: 1099.0 ± 46.3 arcsec). For this, we reduced the square size gradually. We excluded participants whose thresholds were still below the criterion even after reducing the square size to 70% of the original size (120 arcmin). In total, nine from the initial 37 participants were excluded right after the calibration session; thus, the excluded participants did not continue the main experiment. The side length of the squares varied between 84 to 114 arcmin, depending on participants. Accordingly, the square-to-square distance for the experimental conditions with multiple squares varied between 21 to 28.5 arcmin.

Common (Pooled) Conditions. In experiments 1, 2, and 3a (not including Exp. 3b), seven flanker configurations were commonly used for the Vernier discrimination tasks. These flanker configurations were used to test how low-level features interact and their influence on (un)crowding (see Fig. 2). Flanker configurations were arranged vertically or horizontally to test the impact of the radial-tangential anisotropy. In addition, the Vernier targets were in vertical or horizontal orientations to observe the interactions between flanker-target orientations. The configurations were as follows: vernier alone, Vernier with one square, three vertically or horizontally aligned squares, seven vertically or horizontally aligned squares, and 35 (5x7) square grid configurations. For each configuration, the vernier target was either vertically or horizontally oriented. Therefore, overall 14 conditions were tested (Fig. 2). The data from the three experiments were pooled (no participant participated in more than one experiment).

Experiment 1. Eleven participants completed the experiment. We tested the seven aforementioned common configurations and the partial square configurations to investigate possible low-level interactions in uncrowding. As shown in Fig. 3, the partial square configurations had the same number of flanker elements as the common configurations, but only the vertical bars of the squares or only the horizontal bars of the squares. Vernier targets were either vertical or horizontal. Participants were asked to report the Vernier offset direction. For the vertical Vernier, the task was to report whether the lower bar was offset to the left (left button) or right (right button) compared to the upper bar. For the horizontal Vernier, the task was to report whether the right bar was on the top (left button) or bottom (right button) compared to the left bar.

Experiment 2. Ten participants completed the experiment. To test whether uncrowding is universal despite the oblique orientations, in addition to the 14 common conditions (7 flanker configurations x 2 Vernier target orientations), we tested configurations with the flanker configuration tilted by 45˚ and the vernier tilted by either ±45˚ (stimuli details in Fig. 4). For the 45˚ counter-clockwise rotated conditions, the task was to report whether the Vernier bar further away from the fixation dot (outer bar) was offset to the left or right compared to the bar closer to the fixation dot (inner bar). For the 45˚ clockwise rotated conditions, the task was to report whether the inner bar was offset to the top (left) or bottom (right) compared to the outer bar. Each trial started only if the participants kept their eyes fixated on the fixation dot for 150ms.

Experiment 3. In experiment 3a, seven participants completed the experiment. To test whether uncrowding can be explained by crowding of the flanker squares on the center square, a square aspect ratio discrimination task was tested with the seven common configurations, in addition to the Vernier offset discrimination task. For this task, participants were asked to discriminate whether the width or the height of the central square was longer (hence, strictly speaking, the central square was a rectangle). For vertically aligned squares conditions (Fig. 5 a,c,e), the height was adjusted and the width for horizontally aligned squares conditions (Fig. 5 b,d,f,g).

In experiment 3b, 10 participants were tested in the aspect ratio discrimination task as in experiment 3a, but with 2 additional configurations (five vertically or horizontally aligned squares) and with/without the Vernier presentation in the center square. To avoid overlap between the Vernier and the target square, we reduced the Vernier length. Vernier bars were 20 arcmin long, separated by a gap of 2 arcmin. The various conditions are shown in Fig. 5.

Data analysis

We fitted a cumulative Gaussian function (psychometric function) to the data (tested levels and hit rates) and determined the Vernier offset or the square size (Exp. 3) for which 75% correct responses were reached (threshold). Psignifit 3 python toolbox (Fründ et al., 2011) was used for the fitting. High thresholds indicate inferior performance, and low thresholds indicate good performance. Next, we divided the threshold in each condition by the threshold in the Vernier alone condition (threshold elevation). Data were log-transformed to bring the data closer to normality. No obvious violation was detected by visual inspection.

Using R (R Core Team, 2019) and *lme4* package (Bates et al., 2015), we computed linear mixed-effects models (LMM) to account for dependent variables and random variations because of individual differences. The fixed and random effects are specified for each experiment (see *Results* for specifications of each experiment). The model significance (*p*-value) was obtained through likelihood ratio tests (χ^2) by comparing nested models. For each fitted model, using *MuMIn* package (Barton, 2020), we computed the effect size (r^2), that is, the explained variance, when including (conditional r_c^2) and excluding (marginal r_m^2) the random effects (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013). Post-hoc multiple comparisons (Tukey's HSD test) of means were computed with the *multcomp* package (Hothorn et al., 2008).

Model comparisons

We simulated the conditions of experiment 1 (Fig. 3) with Capsule Networks (Doerig et al., 2020, https://github.com/adriendoerig/Capsule-networks-as-recurrent-models-of-grouping-and-segmentation), the Laminart model (Doerig, Bornet, et al., 2019, https://bitbucket.org/albornet/laminart/) and the texture tiling model (TTM; Rosenholtz et al., 2019, https://dspace.mit.edu/handle/1721.1/121152). Capsule networks were trained to recognize Verniers, groups of squares, groups of horizontal bars, and groups of vertical bars presented in isolation (i.e., there were only flankers or the Vernier). After training, the Capsule Network was tested on the different crowding conditions. The model performance was obtained by the percentage of error as in Doerig et al. (Doerig, Schmittwilken, et al., 2020). Performance of the Laminart model was obtained as in Francis et al. (Francis et al., 2017). Performance of TTM was obtained by using an algorithm that matches left and right Vernier templates to mongrels generated by the model (see Bornet et al., Same volume). Model-specific parameters, conditions, and algorithms are available online (https://github.com/Ohyeon5/dissecting_uncrowding).

Results

Pooled conditions. Crowding decreases with the number of squares in vertical and horizontal orientations

In experiments 1, 2, and 3a, we tested the same seven conditions (Vernier alone, 1-square, 3-squares, 7-squares either vertically or horizontally aligned, and 35-squares grid). We pooled the data of 28 participants for these conditions.

We mainly replicated previous findings (Manassi et al., 2013, 2016). When the vertical Vernier was surrounded by a single square, thresholds strongly increased as aimed for. Contrary to previous findings (Manassi et al., 2013, 2016), however, adding only a square both on the left and right of the central square in the horizontally aligned condition did not substantially improve performance (Fig. 2 left.c). Although adding three squares on the left and right (seven squares, horizontally aligned condition) led to a strong decrease of crowding (Fig. 2 left.e), which is in line with previous findings. When squares were vertically aligned (three and seven squares conditions, Fig. 2 left.b & 1 left.d), we also found a decrease of crowding. The same pattern of results holds true when the target Vernier was horizontal (Fig. 2 right). Performance was best with the 35-squares grid (Fig. 2f left&right).

To analyze the relation between threshold elevation and configuration, we computed an LMM with the number of squares in the vertical or horizontal dimension as fixed effects. For example, the horizontally aligned three-squares condition was coded as having three squares in the horizontal dimension and one in the vertical dimension, respectively. Individual participants and target orientations were considered as random intercepts. We found no significant interaction between the two fixed effects (likelihood ratio test between an additive and an interaction model: $\chi^2(1)=0.774$, $p=0.379$). Both fixed effects showed significant differences (horizontal: $\chi^2(1)=60.980$, $p<0.001$; vertical: $\chi^2(1)=36.985$, $p<0.001$). The negative parameter estimates in both dimensions (in Supp. Table 1) show that thresholds significantly decreased when the number of squares increased, which means performance improved. The model explains 38.9% of the variance and only 17.7% when not accounting for the random effects ($r_m^2=0.177$, $r_c^2=0.389$). In addition, we only found a marginal significance of target orientation as a random intercept ($\chi^2(1)=3.853$, $p=0.049$). The difference of explained variance by the models with and without the target orientation as a random intercept is only 1.7% ($r_c^2=0.389$, $r_c^2=0.372$). It seems that qualitative results are similar for both target orientations.

Next, we ran an LMM with only the three and seven flankers configurations to see the possible differences between horizontally and vertically aligned flankers. We included one more fixed effect, namely, flanker orientation (vertical or horizontal); thus, the LMM had three fixed effects and two random intercepts. Interestingly, the fixed effect of the flanker orientations and the number of squares in the horizontal dimension was significant, but not the number of squares in the vertical dimension (flanker orientations: $\chi^2(1)=9.775$, $p<0.01$, horizontal: $\chi^2(1)=21.039$, $p<0.001$, vertical: $\chi^2(1)=0.152$, $p=0.697$). This result indicates that increasing the number of squares in the horizontal dimension improves performance gradually. In contrast, the performance improvement by increasing the number of squares in the vertical dimension was not gradual. Also, post-hoc Tukey's HSD comparison showed that performance improvement with horizontally oriented flankers was significantly better than the vertically oriented ones ($z=3.166$, $p<0.01$). Interactions among fixed effects were not tested because of the dependency. The detailed estimates are reported in Supp. Table 2.

In summary, crowding decreases with the number of squares roughly independent of target and flanker array orientations, despite minor differences.

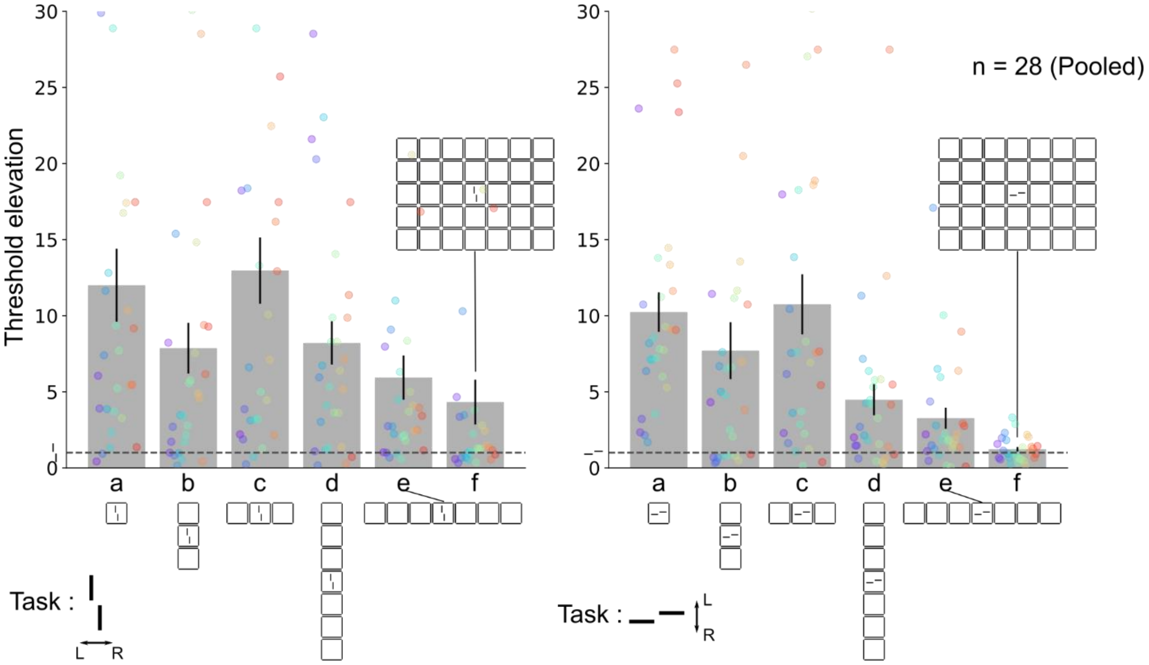


Figure 2. Pooled conditions. The y-axis shows mean threshold elevation (± SEM) relative to the unflanked (Vernier alone) condition (gray dotted lines equal to 1). Larger thresholds represent poor performance (strong crowding), and smaller thresholds represent good performance (weak crowding). Also, performance improves the more squares are presented, independently of the flanker and the Vernier orientation; vertical Vernier left and horizontal Vernier right panel. Colored dots show individual data points.

Experiment 1. Uncrowding cannot be explained by the addition of parts

As shown, regardless of the stimuli orientation, the more flanking squares there are, the better performance is (uncrowding). A major question is to what extent uncrowding depends on low-level interactions (Fig. 1A), such as contour-contour interactions, line-line detector inhibitions, or rather on holistic aspects, such as the good Gestalt of squareness. In other words, can low-level interactions release the strong crowding by the single square around the Vernier target? Here, we systematically dissected the configurations of Fig. 2 in three different ways (Fig. 3): complete square, only the vertical bars of the squares, only the horizontal bars of the squares. Flankers were either horizontally or vertically oriented, and the number of flankers was one, three, to seven flankers. Overall, there were 15 flanker configurations. Note that the central square was always a complete square in the conditions with multiple flankers. The Vernier target was either vertical or horizontal. We call the partial squares whose lines have the same orientation as the target iso-target-flankers and those whose lines are orthogonal ortho-target-flankers.

In the one flanker conditions, performance was worst when the Vernier was flanked by lines of the same orientation, e.g., vertical lines for the vertical Vernier. The LMM was computed with flanker configurations (complete square, iso-, or ortho- target flanker) as a fixed effect and individual participants and target orientations as random intercepts. The fixed effect was significant when compared with an intercept only model (χ^2(2)=19.470, *p*<0.001). Post-hoc Tukey's HSD comparisons indicated that performance with ortho-target-flankers was significantly better than with iso-target flankers (Fig. 3 top&bottom. b vs. c; *z*=4.135, *p*<0.001) and the complete square (Fig. 3 top&bottom.a vs. c; *z*=4.235, *p*<0.001). We found no significant difference between iso-target flankers and the complete squares conditions (Fig. 3 top&bottom a vs. b; *z*=0.099, *p*<0.995). The model explains 34.7% of the variance and 23.5% when not accounting for the random effects (r_m^2=0.235, r_c^2=0.347). The detailed parameter estimates are shown in Supp. Table 3.

In the three and seven flankers conditions, the complete square conditions always showed better performance independently of the number of flankers or flanker orientations. The LMM with a fixed effect of flanker configuration and random intercepts of individual participants and target orientation showed a significant fixed effect (Likelihood ratio test compared with the intercepts only model; χ^2 (2)=49.696, *p*<0.001). Post-hoc Tukey's HSD tests showed a significantly better performance with the complete squares than the other two partial square configurations (complete squares vs. iso -target flankers, *z*=5.060, *p*<0.001; complete squares vs. ortho-target flankers, *z*=7.220, *p*<0.001). Although there appears to be a trend of the flankers to crowd more in the ortho-target flanker conditions than in the iso-target flanker conditions, evidence is not strong enough to make firm claims (Fig. 3; iso- vs. ortho-target flankers, *z*=2.160, *p*=0.078). In addition, even if the effect were significant, the effect size is much smaller than the effect size of crowding vs. uncrowding. This shows that even though iso-target flankers may have a minor influence, it is not the main driving force. The model explains 45.1% of the variance, but only 11.5% when not accounting for the random effects (r_m^2=0.115, r_c^2=0.451). The detailed parameter estimates are shown in Supp. Table 4.

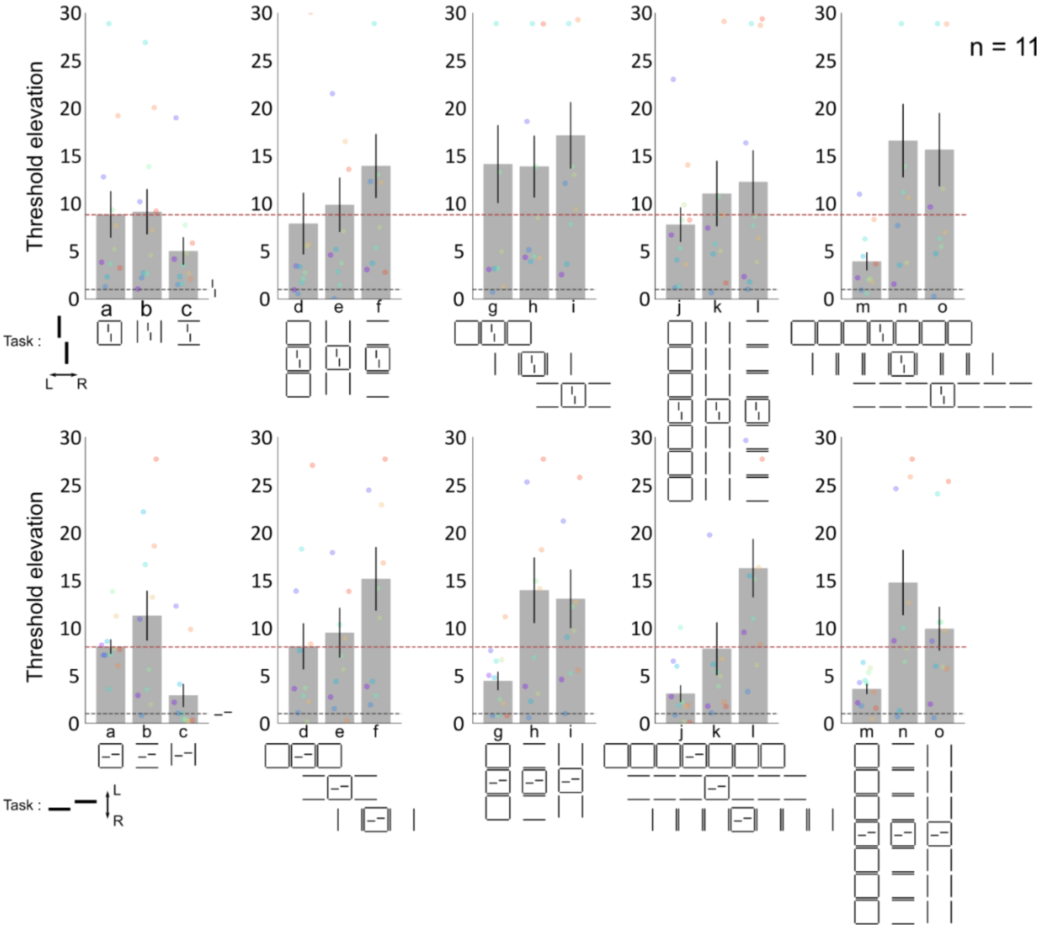


Figure 3. Experiment 1. Systematic dissection of flanker configurations with a vertical (top) or horizontal (bottom) Vernier target. The y-axis shows threshold elevation relative to the unflanked (Vernier alone) condition. In the 1-flanker conditions (**a,b,&c**: crowding conditions), iso-target flankers lead to the same performance deterioration as the complete square (**b vs. a**). In the three- and seven-flankers conditions, complete squares (**d,g,j,&m**) lead to better performance than the iso-target flankers (**e,h,k,&n**) or ortho-target flankers (**f,i,l,&o**). Bars and error bars represent Mean ± SEM, colored dots represent individual data points. Red dotted lines show the performance of the 1 square condition.

Experiment 2. Oblique Orientations

In the pooled conditions, we found no clear differences between vertical and horizontally arranged arrays of squares. Uncrowding seems not to reveal a radial-tangential anisotropy in cardinal orientation, further indicating that low-level aspects, such as the shape of receptive fields in early visual areas, are less important than the overall shape of the configuration. Then what about when a stimulus is presented in oblique orientation (Fig. 1B, 45° arrows)? It is well known that stimuli in cardinal orientations lead to significantly better performance than oblique ones in many visual paradigms (Li et al., 2003; Mach, 1860; Westheimer, 2005) because more neurons are tuned to the cardinal axes (Bauer et al., 1979; Furmanski & Engel, 2000; Xu et al., 2006) or there is an uneven sampling density in the early visual cortex (cortical magnification; Kwon & Liu, 2019; Motter & Simoni, 2007). Oblique orientations may lead to different (un)crowding. There can be three scenarios: (1) a crowding anisotropy between radially (+45°) and tangentially (-45°) arranged array of squares, unlike for cardinal orientation, (2) a similar behavioral pattern as in cardinally oriented stimuli, but with mere performance deterioration, that is, performance improves (uncrowding) as the number of squares increases in either + or – 45 ° direction but not as much as in the cardinal orientation, (3) a completely different behavioral pattern. Here, we tested performance for +45° rotated Verniers in either tangential or radial direction.

Vernier discrimination of the unflanked target was substantially harder in oblique orientations than in the vertical and horizontal orientations (cardinal). Hence, we computed an LMM with stimulus orientations (cardinal or oblique) as the fixed effect and individual participants and the target orientations (vertical, horizontal, -45°, or +45 °) as random intercepts. The fixed effect was significant (likelihood ratio test with the intercept only model; $\chi^2(1)=9.251$, $p<0.01$). Tukey's HSD post-hoc test shows that the performance of the Vernier alone condition with oblique orientations was significantly worse than with cardinal orientations (cardinal vs. oblique; $z=4.753$, $p<0.001$). Detailed estimates are presented in Supp. Table 5.

Contrary to the vertical or horizontal orientations, we did not find a gradual performance improvement as the number of squares increased in one of two orientations. However, there was strong uncrowding in the 35 squares grid configuration, independent of stimulus orientation. An LMM with the number of squares and individual participants as random intercept showed a significant difference for the number of squares (likelihood ratio test with the intercept only model; $\chi^2(1)=32.148$, $p<0.001$). The model explains 26.5% of the variance and 16.6% when not accounting for the random effects ($r_m^2=0.166$, $r_c^2=0.265$). The detailed parameter estimates are presented in Supp. Table 6.

Indeed, it seems that Vernier discrimination is substantially harder in oblique than in cardinal orientations (oblique effect). Also, there is no obvious uncrowding for arrays oriented along the 45° axis. However, for the 35-square grid, neither the oblique orientation of the grid as such nor of the single squares seems to matter. There is clear-cut and strong uncrowding.

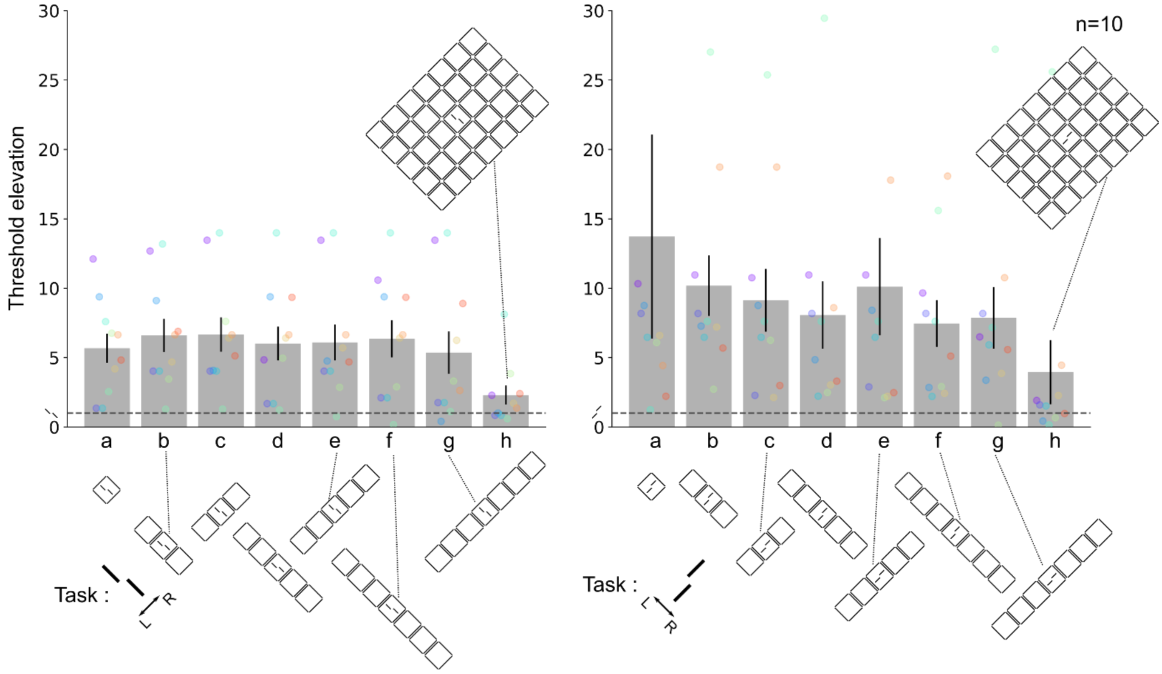


Figure 4. Experiment 2. The left panel shows the -45° rotated Vernier conditions (tangential direction), and right the +45° rotated Vernier conditions (radial direction). The y-axis shows threshold elevation relative to the unflanked (Vernier alone) condition. Performance was poor in most conditions (a-g), regardless of the radial (c,e,g) or tangential (b,d,f) alignments, except with the 35 squares grid (h). Bars and error bars represent Mean ± SEM, colored dots represent individual data points.

Experiment 3. Is uncrowding "crowding of crowding"?

The above experiments showed that uncrowding depends on holistic aspects rather than low-level interactions, regardless of the orientations. It has been suggested crowding is reduced when flankers are suppressed by themselves (Manassi et al., 2013), by flanker awareness (Wallis & Bex, 2011), or by masking

(Chakravarthi & Cavanagh, 2009). Especially, Manassi and colleagues (2013) showed that uncrowding of the Vernier is a consequence of mutual crowding of the squares: Vernier crowding is weak when the central square is crowded by other squares and strong when the square is weakly or not at all crowded. So then, can crowding of crowding fully explain uncrowding?

Consistent with Manassi and colleagues (2013), we found that the aspect ratio of the center square is harder to discriminate as the number of flanking squares increases. Thus the center square was highly crowded by the additional flankers. In addition, we found a crowding anisotropy between horizontally versus vertically aligned squares. The central square was strongly crowded by adding more squares in the horizontal dimension (radial) but not in the vertical dimension (tangential). The LMM was computed with the number of squares in horizontal or vertical dimensions as fixed effects. We coded each flanker as we did in the pooled conditions, that is, three-squares vertically aligned condition as three squares in the vertical dimension and one in the horizontal dimension, respectively. Individual participants and Vernier presentation (experiment 3b only) were considered as random intercepts. LMMs were applied to experiments 3a and 3b separately.

In experiment 3a (Supp. Fig. 1), the two fixed effects had no significant interaction ($\chi^2(1)=1.159$, $p=0.282$). The number of squares in the horizontal dimension had a significant effect ($\chi^2(1)=13.319$, $p<0.001$), whereas the effect in the vertical dimension was not significant ($\chi^2(1)=3.452$, $p=0.063$). In addition, the explained variance difference between the full model with both fixed effects and the nested model without the effect of the vertical dimension was small, only 3.8% (full model: $r_m^2=0.181$, $r_c^2=0.649$, reduced model: $r_m^2=0.143$, $r_c^2=0.602$). Therefore, the number of squares in the vertical dimension may not be a good predictor of the crowding level, whereas the number of squares in the horizontal dimension is a good one. In other words, the number of squares in the horizontal (radial) dimension impacts crowding more than those in the vertical (tangential) dimension, which can be related to the radial-tangential anisotropy of the crowding. The detailed parameter estimates are presented in Supp. Table 7.

The same results hold for experiment 3b (Fig. 5). Two fixed effects showed no significant interaction ($\chi^2(1)=12.682$, $p=0.102$). The number of squares in the horizontal dimension was significant ($\chi^2(1)=27.387$, $p<0.001$), but not for the vertical orientation ($\chi^2(1)=0.116$, $p=0.733$). Again, the explained variance difference was tiny. The difference between the full model, including both effects and the reduced model excluding the effect in the vertical dimension, was only 0.05% (full model: $r_m^2=0.163$, $r_c^2=0.423$, reduced model: $r_m^2=0.163$, $r_c^2=0.423$). The detailed parameter estimates are shown in Supp. Table 8.

To explicitly test that crowding was stronger for horizontally aligned squares than for vertically aligned squares, we computed another LMM. Here, we used the 3, 5, or 7 square conditions only and considered the number of squares and the flanker alignment orientation as fixed effects and the same random intercepts as the tested model. The two fixed effects showed no significant interactions ($\chi^2(1)=3.264$ $p=0.071$). The flanker alignment orientation showed a significant effect ($\chi^2(1)=17.848$, $p<0.001$). The number of squares had no significant effect ($\chi^2(1)=0.122$, $p=0.727$). Although the interaction model showed no significant effect, there was a trend for an interaction, that is, crowding increased with the number of squares in the horizontal but not clear in the vertical orientation. However, the interaction is minor compared to the effect of the flanker alignment orientation. This minor interaction may be a reason why the fixed effect of the number of squares did not show significance. Post-hoc Tukey's HSD tests showed a significantly stronger crowding for the horizontally aligned squares than for the vertically aligned squares (horizontal vs. vertical; $z=4.404$, $p<0.001$). The detailed parameter estimates are shown in Supp. Table 9. The results are consistent with the well-known crowding radial-tangential anisotropy (Toet & Levi, 1992).

In addition, the presentation of a Vernier in the central square does not affect performance (experiment 3b, Fig. 5 left vs. right), i.e., crowding was not due to target location uncertainty. We used the LMM with two fixed effects, namely, the number of squares in each dimension and random intercepts for individual participants and Vernier presentation (the same LMM as applied to experiment 3b, Supp. Table 8). The likelihood ratio test showed no significant difference between the full model and the model excluding Vernier presentation ($\chi^2(1)=0.167$, $p=0.683$). Also, the explained variance difference between both models was little, only 0.5% (full model: $r_m^2=0.163$, $r_c^2=0.423$, reduced model: $r_m^2=0.164$, $r_c^2=0.418$).

In summary, flankers aligned in the horizontal (radial) dimension crowd stronger than in the vertical (tangential) dimension. However, such an anisotropy was not reflected in the Vernier discrimination task; that is, the Vernier performance was not better in horizontally aligned squares than in vertically aligned squares (Fig. 2, further discussion in *Discussion*).

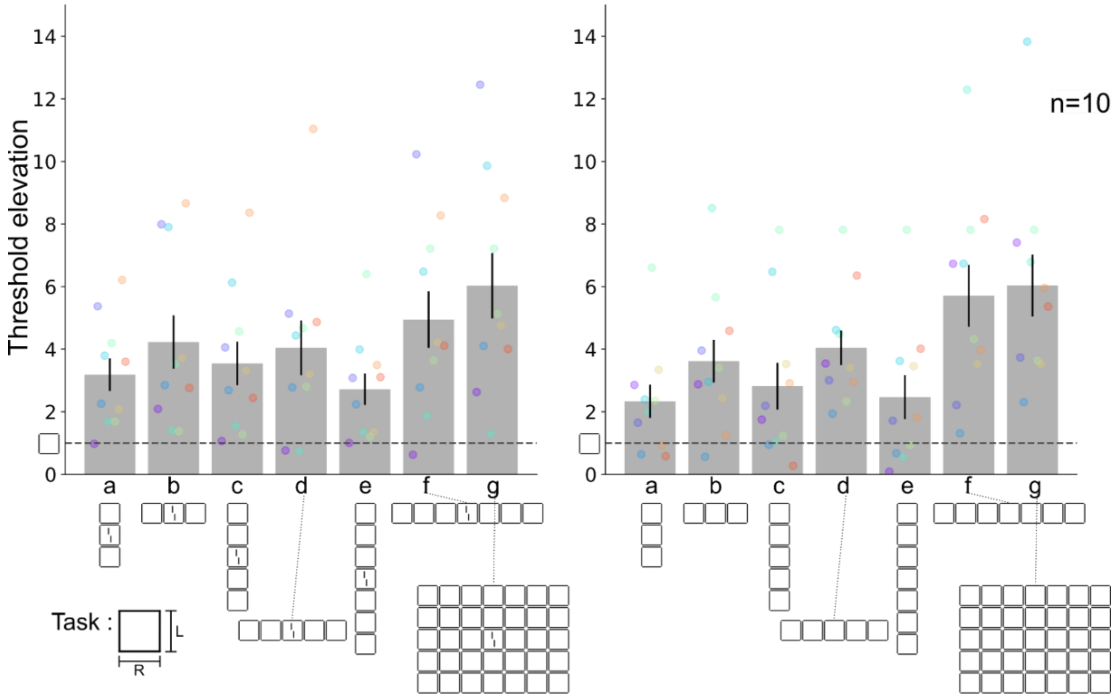


Figure 5. Experiment 3b. The center square aspect ratio discrimination task with (left) and without (right) Vernier presentation. Performance deteriorated (the target was more crowded) as the number of squares increased in the horizontal dimension, independent of whether or not the Vernier was presented. The y-axis shows threshold elevation relative to the one square condition. Mean ± SEM, colored dots represent individual data points. Note the change of y-axis scaling.

Models. Model comparison suggests that object-based grouping is needed to explain uncrowding

In the above experiments, we showed that uncrowding happens regardless of orientation and depends on holistic, rather than local, aspects of the stimulus. Here, we tested three models, which take global aspects into account but are based on different premises. Capsule networks and the Laminart model are two-stage models, in which elements are first parsed into different groups, and then interference occurs only within the groups. Capsule networks group elements on the basis of object-level routing by agreement (for details, see Doerig et al., 2020; Sabour et al., 2017), whereas the Laminart model groups elements on the basis of low-level features (for details, see Francis et al., 2017; Bornet et al., 2019). The TTM model is a one-stage model that pools many low-level features computed over pooling regions whose size grows with eccentricity (for details, see Rosenholtz et al., 2019). We tested the vertical Vernier target conditions of experiment 1. Here, we only show results obtained with the horizontally aligned flanker conditions (Fig. 6). The model results for the vertically aligned flanker conditions are comparable (Supp. Fig. 4).

Capsule Network reproduced the general human behavior pattern well, that is, performance improved when adding more squares (Fig. 6Ad, 6Ag; red bars) and deteriorated when adding either the iso- or ortho-target flankers (Fig. 6Ae, 6Af, 6Ah, 6Ai; gray bars). Note that there were still minor performance differences, for example, human performance for only vertical lines (Fig. 6Eb) was equally bad as in the one-square condition, but the model performance was much better (Fig. 6Db vs. Fig 6Ab). The Laminart model partially reproduced the human behavior, that is, performance improved when adding more squares (Fig. 6Bd, 6Bg; red bars) and deteriorated when adding the iso-target flankers (Fig. 6Be, 6Bh). However, unlike humans, model performance improved when adding ortho-target flankers (Fig. 6B f,i). In both models, the performance of the complete square conditions could not be explained by simply adding the performances of the iso- and ortho-target flankers conditions, that is, the performance of Fig. 6Ad was smaller than Fig. 6Ae or Fig. 6Af.

The TTM (1-stage model) could not reproduce the human behavior, that is, adding more squares deteriorated performance, and performances with iso- or ortho-target flankers were better than in the complete squares conditions (Fig. 6C). Moreover, performance in the complete squares conditions could more or less be explained by adding the performance levels of the corresponding iso- and ortho-target flankers conditions (i.e., the performance of Fig. 6C d was roughly equal to Fig. 6C e plus Fig. 6C f).

In addition, we trained three control networks using the exact same training procedure as we used for the Capsule networks: a feedforward CNN and two recurrent CNNs. These networks had the same number of layers and neurons, and the only differences were in the connectivity between the neurons in the last two layers. This allowed us to control: (1) whether the training regime is sufficient to explain the experimental results, even without recurrent grouping and segmentation, and (2) whether any kind of recurrence is sufficient vs. whether specific grouping and segmentation processing of Capsule networks is needed. The results clearly show that these control networks do not reproduce our results, supporting our claim that grouping and segmentation processes are needed (Supp. Fig. 2) for uncrowding (Fig. 3). Moreover, the Laminart model and TTM were tested with different parameters. Changes in the model parameters did not lead to obvious differences (Supp. Fig. 3). Hence, in summary, our results favor the two-stage models over the one-stage model.

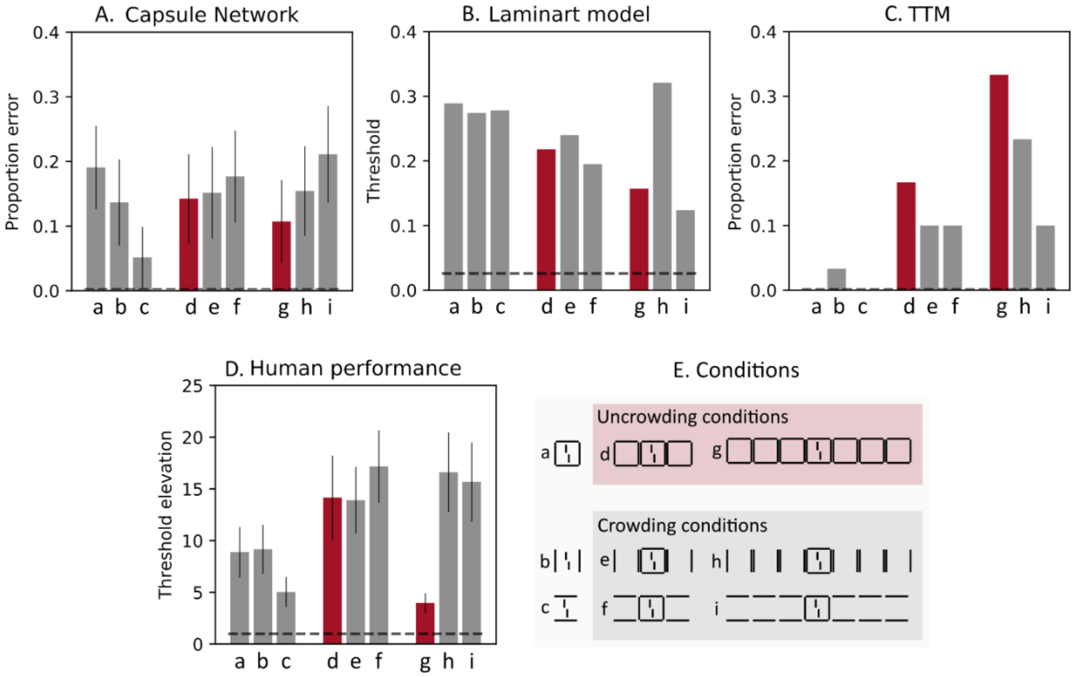


Figure 6. Model performance: percent error for Capsule networks and TTM, Vernier offset thresholds for the Laminart model. For both measures, larger values indicate worse performances. Red bars represent conditions leading to uncrowding in humans (good performance), and gray bars represent crowding (poor performance). Gray dashed lines show the model performance for the Vernier only condition. (A) Performance of Capsule Networks. We averaged the proportion of errors from 10 separately trained networks (mean±SEM). (B) Performance of the Laminart model. We used an inference mechanism as described in Francis et al., 2017, and averaged the results over 20 runs per condition. (C) Performance of the TTM. We created 15 mongrels per condition and per offset direction (in total, 30 mongrels per condition) and determined the proportion of errors using a template matching algorithm. (D) Human performance reordered from Fig 3. (E) Conditions tested. Vertically aligned flanker conditions were also tested and presented in Supp. Fig. 4.

Discussion

Crowding is at the heart of vision research as elements are rarely encountered in isolation. However, even after a century of research (e.g., Korte, 1923; Ehlers, 1936; Flom et al., 1963; Bouma, 1970), the mechanisms underlying crowding are still largely unknown and controversially discussed. Classically, crowding was explained by local models, where only *neighboring* elements with *similar* features interact with each other, for example, via lateral inhibition (Carandini & Heeger, 2012). Alternatively, the outputs of the neurons may be pooled (Dakin et al., 2010; Greenwood et al., 2009, 2017; Parkes et al., 2001; Pelli, 2008; Rosenholtz et al., 2012, 2019), features may be substituted (Huckauf & Heller, 2002; Strasburger, 2005; Strasburger et al., 1991), or crowding may be mediated by top-down processes (He et al., 1996; Montaser-Kouhsari & Rajimehr, 2005; Tripathy & Cavanagh, 2002; Yeshurun & Rashal, 2010). Such models were motivated by experiments showing that, for example, crowding strongly decreases when target and flanker have different contrast polarity (Kooi et al., 1994), color (Kennedy & Whitaker, 2010; van den Berg et al., 2007), motion (Bex & Dakin, 2005), and more. Likewise, crowding decreases when flankers are moved away from the target, which is often described by Bouma's law stating that flankers interfere only within a window of half the eccentricity around the target (Bouma, 1973).

However, all these explanations fall apart when more flankers are presented. Flankers outside Bouma's window can suppress crowding up to the performance level of the unflanked target (Fig. 2f, grid condition). Similar effects have been shown previously with various stimuli such as Verniers (Manassi et al., 2012, 2013, 2015, 2016; Sayim et al., 2010), Gabors (Levi & Carney, 2009; Livne & Sagi, 2007; Maus et al., 2011; Saarela et al., 2009; Saarela & Herzog, 2008), shapes (Kimchi & Pirkner, 2015), letters (Reuther & Chakravarthi, 2014; Saarela et al., 2010), textures (Herrera-Esposito et al., 2020), as well as in haptics (Overvliet & Sayim, 2016) and audition (Oberfeld & Stahn, 2012). Feature similarity is important but not decisive because strong crowding can also occur with flankers of different contrast polarity and color (Manassi et al., 2012; Sayim et al., 2008). What matters is the configuration (Livne & Sagi, 2007) of potentially all elements across large parts of the visual field (Herzog et al., 2016; Herzog & Manassi, 2015). For this reason, simple local (pooling) models have been largely but not fully abandoned. For example, Greenwood and colleagues (2020) take configuration effects in crowding as "modulations" of a local pooling mechanism. However, we think that strong effects such as uncrowding, going from good performance for a single target to strong crowding with a single square (threshold elevation of 12, Fig. 2) to uncrowding with the 35 squares (threshold elevation of 4, Fig. 2), are beyond what can be called a modulation.

Here, we have tested to what extent the overall configuration plays a role in crowding by dissecting uncrowding configurations systematically. First, we reproduced previous findings of uncrowding with an increasing number of elements. Performances in the 35 square grid condition were about at the same level as in the Vernier alone condition (Fig. 2f). Importantly, despite minor differences, uncrowding occurs for both the horizontal and vertical arranged flankers, also for horizontal and vertical Vernier targets (Figs. 2 and 3). Note that, unlike previous work (Manassi et al., 2013, 2016), the three horizontally aligned squares did not show a clear performance improvement compared to the single square condition. We do not have an explanation for this beyond noises. The oblique configuration showed a different behavioral pattern (Fig. 4). For example, increasing the number of squares in one of the ±45° orientations did not improve performance (Fig. 4, conditions a vs. c, e, and g, or conditions a vs. b, d, and f). Surprisingly, when the entire 35-square grid was presented, the performance was as good as for the cardinal orientations (Fig. 4h left & right, 35-square grid condition). Livne and Sagi (2011) showed that obliquely oriented and positioned flankers crowd stronger than cardinally oriented flankers. In addition, the obliquely presented stimuli made various visual tasks significantly harder, including orientation discrimination (Bouma & Andriessen, 1968), orientation discrimination under crowding (Livne & Sagi, 2011), Vernier discrimination (Saarinen & Levi, 1995; Westheimer, 2005), motion discrimination (Ball & Sekuler, 1982; Coletta et al., 1993), orientation detection (Attneave & Olson, 1967), and more, likely because of neuronal preferences of cardinal orientations in low-level visual areas (Bauer et al., 1979; Furmanski & Engel, 2000; Li et al., 2003; Xu et al., 2006). Hence, similarly, we expected performance would deteriorate with the oblique flankers while keeping crowding characteristics similar to the cardinal flankers. However, the results were not as expected. There was no performance improvement (uncrowding) when increasing the number of squares in either ± 45 orientations but only with the 35-square grid (Fig. 4). This may be because the grouping cue was too weak with three, five, or seven squares for one of the ±45° orientations. Nevertheless, when a stronger grouping cue was provided by the grid of 35 squares, performance was good (uncrowding) regardless of orientation, approaching the performance in Vernier alone conditions (and comparable to the cardinal stimuli's performance). Therefore, our results once again argue for complex spatial interactions, which most existing models cannot capture easily.

Second, crowding and uncrowding in the multi-square conditions cannot be explained by local interactions of its subparts—the configuration matters. Alternatively, local interactions were important for the single flanker conditions (Fig. 3a, 3b, & 3c, upper & bottom panels; Supp. Table 3). The LMM and post-hoc comparisons showed that ortho-target flanker conditions had significantly better performance, whereas iso-target flanker conditions had comparable performances to complete square conditions. More specifically, performance for the vertical Vernier surrounded by the square is as poor as for the vertical lines of the square only (Fig. 3 upper panel a & b). This result may be taken as support that only neurons of similar orientation interact. However, there is still some effect of the horizontal lines too, which may be considered an unspecific effect (Fig. 3 upper panel c). This effect is even more pronounced for the horizontal Vernier since the horizontal lines crowd more than the square (Fig. 3 lower panel a & b). On the other hand, (un)crowding in the multi-square conditions showed clearly a different pattern (Fig 3d-3o upper & bottom panels; Supp. Table 4). In general, conditions with complete squares lead to better or equal performance than conditions with parts of a square only, except for an iso-target condition (Fig. 3 h, upper panel), indicating good Gestalt matters. In the three and seven squares conditions, post-hoc Tukey's HSD test after an LMM analysis showed that complete squares flanker configurations led to significantly better performance than iso-target and ortho-target flanker configurations. Therefore, our results imply that, unlike complete squares, parts of the squares cannot release crowding by low-level interactions, such as contour-contour interaction or line-line detector inhibition.

Third, as a side note, the results in Fig. 3 conditions b and c also show that participants did not perceive the task as a bisection task. In other words, participants did not discriminate the Vernier offset relative to the bisector (bisection cue) of two parallel bars, as in a bisection task (e.g., Clarke et al., 2014). Since performance with iso-target flankers was worse in this condition than with the ortho-target flankers, no bisection cue can be used (to be more precise: if there were a bisection cue, it must be much weaker than other mechanisms involved).

Fourth, there were complex interactions between Vernier orientation and square configuration orientation, which cannot easily be explained by a single, local mechanism, except for the single iso- and ortho-target flankers, which are in accordance with the predictions of most local models. However, for the more complex configurations, Vernier orientation did not matter. For example, performance for the vertical and horizontal Vernier showed a very similar pattern independent of Vernier orientation for the horizontally arranged squares: strong crowding for one and three squares and strong improvement of the seven squares conditions. The random intercept of target orientation only had a marginal significance; also, the explained variance with and without the random intercept was only 1.7% (in Pooled conditions). However, qualitatively, there was a trend of an effect of the vertically oriented square array with Vernier orientation. There is only a weak improvement, if at all, for the vertical Vernier as the number of squares increases (Fig. 2 left b vs. d, weaker in Fig. 3 upper d vs. j), but there is a clear improvement for the horizontal target (Fig. 2 right b vs. d, weaker in Fig. 3 lower g vs. m). Again, this latter effect cannot be explained by an increase in the number of horizontal lines because performance deteriorates the more horizontal lines there are. Thus mutual inhibition between the horizontal lines is not a viable explanation. Finally, rotating the entire configuration showed a different behavioral pattern, that is, increasing the number of squares only in one orientation (either ±45°) did not improve performance, but significant performance improvements were observed with 35-square grid conditions. However, again, Vernier orientation did not matter.

Fifth, we reproduced the previous finding that the mutual crowding of the squares increases with the number of flanking squares (Manassi et al., 2013). In addition, we found a radial-tangential anisotropy (Chung, 2013; Greenwood et al., 2017; Kwon et al., 2014; Malania et al., 2020; Toet & Levi, 1992). The target square in the horizontally aligned squares was more crowded than in the vertically aligned squares (Fig. 5 & Supp Fig. 1). However, such anisotropy is not reflected in the Vernier discrimination task, that is, the Vernier performance was not better in horizontally aligned squares than in vertically aligned squares (Fig. 2). If uncrowding can be simply explained by "crowding of crowding" as Manassi and colleagues (2013) suggested, stronger crowding in horizontally aligned squares would have induced better segregation of squares from Vernier target, hence, better performance with horizontally aligned squares than with vertically aligned squares. However, this was not the case. In the 3-squares conditions, to the contrary, vertically aligned squares led to better performance than the horizontally aligned squares (Fig. 2b vs. 2c), but not in the seven-squares conditions (Fig. 2d vs. 2e). The results again suggest that uncrowding is not a single process but rather a complex problem with many factors involved.

Sixth, we found no significant differences with and without the Vernier stimulus in the center square indicating the target position (the explained variance difference between with and without the random intercept of Vernier presentation was small, only 0.5%; Fig. 5 left vs. right). This result indicates that performance deterioration does not come from location uncertainty.

In the current work, we did not compute statistics for all possible comparisons between conditions in experiments to avoid multiple testing and because they were not part of our main research question. The majority of the analyzed comparisons show that the holistic structure matters (e.g., Exp. 1, complete square conditions vs. partial square conditions).

Whereas certain types of element-element interactions might explain single conditions, it seems that the entirety of findings resists such an explanation. Likely, there are many mechanisms in operation, and these mechanisms may be found more on an implicit statistical level than by explicit element-element interactions similar to the processing of CNNs where single neurons code for a large number of stimulus features. For this reason, we subjected our data to two 2-stage models, which take large-scale configural interactions into account (Laminart and Capsule networks), and a 1-stage model, which was proposed to account for complex configurational effects with high-dimensional pooling (HD pooling) and in the decision process. Other models, such as classic CNNs, epitomes, Fourier analysis, etc., failed with the basic crowding conditions and were not considered here (Doerig et al., 2019).

The TTM did not show uncrowding when adding more squares. Albeit its ability to pool a large number of features (HD pooling), the information of the target Vernier and the precise flanker structure was irretrievably lost. Whereas TTM is an excellent model for textural processing and summary statistics, we suggest that TTM misses a flexible segmentation stage, which segments visual scenes in multiple groups depending on the configuration. The TTM, as a 1-stage model, does not have a flexible segmentation stage and thus treats fine details of all elements equally. For this reason, it erases small details, which makes a major difference for the human system and leads to qualitatively different results (Wallis et al., 2019). In addition, there is a similar problem with the pooling regions. As shown in the experiments, changes across large parts of the visual field matter. For example, the outmost squares strongly matter but are 8.5° away from the Vernier target. Using wider filters to take this information into account would strongly compress the target. Hence, further detailed information is crucial, thus, more flexible architectures are needed.

The Laminart model reproduced human performance when more square flankers were added (uncrowding) but unexpectedly showed uncrowding in the iso-target conditions (Fig. 6B f&i), whereas human participants showed strong crowding (Fig. 6D f&i). Capsule networks reproduced the results best as they take explicit object representations into account, suggesting that object-level segmentation is needed to fully account for the complex effects of configuration. However, Capsule networks were trained for the specific stimuli and task, whereas TTM and Laminart were not adapted. Nonetheless, the human-like performance of Capsule networks was not due to the training process, since the control networks, without grouping and segmentation process, using the same training procedure, could not reproduce the human performance

(Supp. Fig. 2). Thus, these results support that object-based grouping and segmentation processes are crucial to explain human behavior.

We believe that our results show that flexible segmentation and grouping are critical for human vision (as do Capsule networks and Laminart model). In natural conditions, nearby elements on the retina may not be nearby in the outer world because they may be located at very different depth planes (perceptual groups). For example, a mesh fence in front of a house leads to overlapping contours of the fence and the house in the early visual areas. A flexible grouping and segmentation stage first groups these contours with each other before any interaction occurs across the depth planes. Crowding occurs when the individual contours within the depth plane may be suppressed to see the wholes, such as the fence and the house. No crowding should occur between contours that do not belong to the same depth plane. Indeed, crowding does not occur when the target and the flankers belong to different depth planes, even though they lie at nearby locations in retinal coordinates (Astle et al., 2014; Kooi et al., 1994; Sayim et al., 2008). In our experiments, a single square and Vernier are grouped as one object, that is, they belong to a single depth plane. In contrast, with the number of squares, square flankers are grouped, and the Vernier is assigned to a different group, either because it is perceived as belonging to a different depth plane or to a different object in the same depth plane.

In summary, we are still quite far from understanding and explaining the major characteristics of crowding. A model that can explain the major characteristics of crowding, in a nutshell, does not exist yet. We are, however, optimistic that such a model exists, since crowding shows universal characteristics across all types of stimuli (Herrera-Esposito et al., 2020; Herzog et al., 2015; Kimchi & Pirkner, 2015; Levi & Carney, 2009; Pelli et al., 2004; Reuther & Chakravarthi, 2014; Saarela et al., 2009; van den Berg et al., 2007; Wallace & Tjan, 2011), tasks (Farzin et al., 2009; Fischer & Whitney, 2011; Yeh et al., 2012), and modalities (Oberfeld & Stahn, 2012; Overvliet & Sayim, 2016). Understanding crowding may unearth the strategies that are used to make sense of the outer world.

Acknowledgment

We thank JAFR for collecting the data. OHC, MHH were supported by the Swiss National Science Foundation (SNF) 320030_176153 “Basics of visual processing: from elements to figures”. AB was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 785907 (Human Brain Project SGA2) and No. 945539 (Human Brain Project SGA3). AD was supported by the SNF postdoc grant N.191718 “Towards machines that see like us: human eye movements for robust deep recurrent neural networks”.

References

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

Astle, A. T., McGovern, D. P., & McGraw, P. V. (2014). Characterizing the role of disparity information in alleviating visual crowding. *Journal of Vision*, *14*(6), 8–8. https://doi.org/10.1167/14.6.8

Attneave, F., & Olson, R. K. (1967). Discriminability of stimuli varying in physical and retinal orientation. *Journal of Experimental Psychology*, *74*(2p1), 149.

Bach, M. (1996). The Freiburg Visual Acuity Test—Automatic Measurement of Visual Acuity. *Optometry and Vision Science*, *73*(1), 49–53.

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science*, *218*(4573), 697–698.

Barton, K. (2020). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1--48. https://doi.org/10.18637/jss.v067.i01

Bauer, J. A., Owens, D. A., Thomas, J., & Held, R. (1979). Monkeys show an oblique effect. *Perception*, *8*(3), 247–253.

Bex, P. J., & Dakin, S. C. (2005). Spatial interference among moving targets. *Vision Research*, *45*(11), 1385–1398.

Bornet, A., Choung, O.-H., Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (Under review). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Journal of Vision*.

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.

Bouma, H. (1970). Interaction Effects in Parafoveal Letter Recognition. *Nature*, *226*(5241), 177–178. https://doi.org/10.1038/226177a0

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research*, *13*(4), 767–782.

Bouma, H., & Andriessen, J. J. (1968). Perceived orientation of isolated line segments. *Vision Research*, *8*(5), 493–507.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Chakravarthi, R., & Cavanagh, P. (2009). Recovery of a crowded object by masking the flankers: Determining the locus of feature integration. *Journal of Vision*, *9*(10), 4–4. https://doi.org/10.1167/9.10.4

Chicherov, V., & Herzog, M. H. (2015). Targets but not flankers are suppressed in crowding as revealed by EEG frequency tagging. *NeuroImage*, *119*, 325–331. https://doi.org/10.1016/j.neuroimage.2015.06.047

Chicherov, V., Plomp, G., & Herzog, M. H. (2014). Neural correlates of visual crowding. *NeuroImage*, *93*, 23–31. https://doi.org/10.1016/j.neuroimage.2014.02.021

Chung, S. T. L. (2013). Cortical Reorganization after Long-Term Adaptation to Retinal Lesions in Humans. *Journal of Neuroscience*, *33*(46), 18080–18086. https://doi.org/10.1523/JNEUROSCI.2764-13.2013

Chung, S. T. L., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research*, *41*(14), 1833–1850. https://doi.org/10.1016/S0042-6989(01)00071-2

Clarke, A. M., Grzeczkowski, L., Mast, F. W., Gauthier, I., & Herzog, M. H. (2014). Deleterious effects of roving on learned tasks. *Vision Research*, *99*, 88–92. https://doi.org/10.1016/j.visres.2013.12.010

Clarke, A. M., Herzog, M. H., & Francis, G. (2014). Visual crowding illustrates the inadequacy of local vs. Global and feedforward vs. Feedback distinctions in modeling visual perception. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01193

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, *18*, 1648.

Coletta, N. J., Segu, P., & Tiana, C. L. (1993). An oblique effect in parafoveal motion perception. *Vision Research*, *33*(18), 2747–2756.

Dakin, S. C., Cass, J., Greenwood, J. A., & Bex, P. J. (2010). Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli. *Journal of Vision*, *10*(10), 14–14.

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. https://doi.org/10.1371/journal.pcbi.1006580

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.

Ehlers, H. (1936). V: The movements of the eyes during reading. *Acta Ophthalmologica*, *14*(1-2), 56–63.

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision*, *9*(6), 18.1-15. https://doi.org/10.1167/9.6.18

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, *106*(3), 1389–1398.

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), 483–504. https://doi.org/10.1037/rev0000070

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6), 16–16. https://doi.org/10.1167/11.6.16

Furmanski, C. S., & Engel, S. A. (2000). An oblique effect in human primary visual cortex. *Nature Neuroscience*, *3*(6), 535–536.

Gheri, C., Morgan, M. J., & Solomon, J. A. (2007). The relationship between search efficiency and crowding. *Perception*, *36*(12), 1779–1787.

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.

Greenwood, J. A., & Parsons, M. J. (2020). Dissociable effects of visual crowding on the perception of color and motion. *Proceedings of the National Academy of Sciences*, *117*(14), 8196–8202.

Greenwood, J. A., Szinte, M., Sayim, B., & Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceedings of the National Academy of Sciences*, *114*(17), E3573–E3582. https://doi.org/10.1073/pnas.1615504114

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, *383*(6598), 334–337. https://doi.org/10.1038/383334a0

Herrera-Esposito, D., Coen-Cagli, R., & Gomez-Sena, L. (2020). Flexible contextual modulation of naturalistic texture perception in peripheral vision. *BioRxiv*, 2020.01.24.918813. https://doi.org/10.1101/2020.01.24.918813

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.

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. https://doi.org/10.1167/15.6.5

Herzog, M. H., Thunell, E., & Ögmen, H. (2016). Putting low-level vision into global context: Why vision cannot be reduced to basic circuits. *Vision Research*, *126*, 9–18. https://doi.org/10.1016/j.visres.2015.09.009

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Hubel, D. H., Wiesel, T. N., & Stryker, M. P. (1978). Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, *177*(3), 361–379.

Huckauf, A., & Heller, D. (2002). What various kinds of errors tell us about lateral masking effects. *Visual Cognition*, *9*(7), 889–910.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Kennedy, G. J., & Whitaker, D. (2010). The chromatic selectivity of visual crowding. *Journal of Vision*, *10*(6), 15–15.

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.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

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

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

Kwon, M., Bao, P., Millin, R., & Tjan, B. S. (2014). Radial-tangential anisotropy of crowding in the early visual areas. *Journal of Neurophysiology*, *112*(10), 2413–2422. https://doi.org/10.1152/jn.00476.2014

Kwon, M., & Liu, R. (2019). Linkage between retinal ganglion cell density and the nonuniform spatial integration across the visual field. *Proceedings of the National Academy of Sciences*, *116*(9), 3827–3836. https://doi.org/10.1073/pnas.1817076116

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. https://doi.org/10.1016/j.visres.2007.12.009

Levi, D. M., & Carney, T. (2009). Crowding in peripheral vision: Why bigger is better. *Current Biology*, *19*(23), 1988–1993.

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: A neural basis in the visual cortex. *Journal of Neurophysiology*, *90*(1), 204–217.

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision*, *7*(2), 4. https://doi.org/10.1167/7.2.4

Livne, T., & Sagi, D. (2011). Multiple levels of orientation anisotropy in crowding with Gabor flankers. *Journal of Vision*, *11*(13), 18–18. https://doi.org/10.1167/11.13.18

Mach, E. (1860). Ueber das Sehen von Lagen und Winkeln durch die Bewegung des Auges. *Sitzungsberichte Der Math Cl Der Kais Akad Der Wissenschaften*, *42*, 215–224.

Malania, M., Herzog, M. H., & Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *Journal of Vision*, *7*(2), 1–1. https://doi.org/10.1167/7.2.1

Malania, M., Pawellek, M., Plank, T., & Greenlee, M. W. (2020). Training-Induced Changes in Radial–Tangential Anisotropy of Visual Crowding. *Translational Vision Science & Technology*, *9*(9), 25–25. https://doi.org/10.1167/tvst.9.9.25

Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision*, *15*(8), 16–16. https://doi.org/10.1167/15.8.16

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

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. https://doi.org/10.1167/12.10.13

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision*, *13*(13), 10–10. https://doi.org/10.1167/13.13.10

Maus, G. W., Fischer, J., & Whitney, D. (2011). Perceived Positions Determine Crowding. *PLOS ONE*, *6*(5), e19796. https://doi.org/10.1371/journal.pone.0019796

Miles, W. R. (1930). Ocular dominance in human adults. *The Journal of General Psychology*, *3*(3), 412–430.

Montaser-Kouhsari, L., & Rajimehr, R. (2005). Subliminal attentional modulation in crowding condition. *Vision Research*, *45*(7), 839–844. https://doi.org/10.1016/j.visres.2004.10.020

Motter, B. C., & Simoni, D. A. (2007). The roles of cortical image separation and size in active visual search performance. *Journal of Vision*, *7*(2), 6–6. https://doi.org/10.1167/7.2.6

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (n.d.). *The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded*. 11.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Oberfeld, D., & Stahn, P. (2012). Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PloS One*, *7*(10), e48054.

Overvliet, K. E., & Sayim, B. (2016). Perceptual grouping determines haptic contextual modulation. *Vision Research*, *126*, 52–58. https://doi.org/10.1016/j.visres.2015.04.016

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. https://doi.org/10.1038/89532

Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology*, *18*(4), 445–451.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12), 12. https://doi.org/10.1167/4.12.12

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135. https://doi.org/10.1038/nn.2187

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

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

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

R Core Team. (2019). *R: A Language and Environment for Statistical Computing* (Vienna, Austria). R Foundation for Statistical Computing. https://www.R-project.org/

Reuther, J., & Chakravarthi, R. (2014). Categorical membership modulates crowding: Evidence from characters. *Journal of Vision*, *14*(6), 5–5. https://doi.org/10.1167/14.6.5

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.

Rosenholtz, R., Yu, D., & Keshvari, S. (2019). Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of Vision*, *19*(7), 15–15. https://doi.org/10.1167/19.7.15

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. https://doi.org/10.1167/8.3.23

Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, *9*(2), 5–5. https://doi.org/10.1167/9.2.5

Saarela, T. P., Westheimer, G., & Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *Journal of Vision*, *10*(10), 17–17. https://doi.org/10.1167/10.10.17

Saarinen, J., & Levi, D. M. (1995). Orientation anisotropy in vernier acuity. *Vision Research*, *35*(17), 2449–2461.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 3856–3866.

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in vernier acuity. *Journal of Vision*, *8*(8), 12–12. https://doi.org/10.1167/8.8.12

Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt Factors Modulate Basic Spatial Vision. *Psychological Science*, *21*(5), 641–644. https://doi.org/10.1177/0956797610368811

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

Silson, E. H., Reynolds, R. C., Kravitz, D. J., & Baker, C. I. (2018). Differential Sampling of Visual Space in Ventral and Dorsal Early Visual Cortex. *The Journal of Neuroscience*, *38*(9), 2294–2303. https://doi.org/10.1523/JNEUROSCI.2717-17.2018

Solomon, J. A., Felisberti, F. M., & Morgan, M. J. (2004). Crowding and the tilt illusion: Toward a unified account. *Journal of Vision*, *4*(6), 9–9. https://doi.org/10.1167/4.6.9

Strasburger, H. (2005). Unfocussed spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision*, *5*(11), 8–8. https://doi.org/10.1167/5.11.8

Strasburger, H. (2020). Seven Myths on Crowding and Peripheral Vision. *I-Perception*, *11*(3), 2041669520913052. https://doi.org/10.1177/2041669520913052

Strasburger, H., Harvey, L. O., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Perception & Psychophysics*, *49*(6), 495–508.

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient Estimates on Probability Functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787. https://doi.org/10.1121/1.1910407

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357. https://doi.org/10.1016/0042-6989(92)90227-A

Tripathy, S. P., & Cavanagh, P. (2002). The extent of crowding in peripheral vision does not scale with target size. *Vision Research*, *42*(20), 2357–2369. https://doi.org/10.1016/S0042-6989(02)00197-9

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, *7*(2), 14–14.

van der Burg, E., Olivers, C. N. L., & Cass, J. (2017). Evolving the keys to visual crowding. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(4), 690–699. https://doi.org/10.1037/xhp0000337

Wallace, J. M., & Tjan, B. S. (2011). Object crowding. *Journal of Vision*, *11*(6), 19–19. https://doi.org/10.1167/11.6.19

Wallis, T. S. A., & Bex, P. J. (2011). Visual crowding is correlated with awareness. *Current Biology : CB*, *21*(3), 254–258. https://doi.org/10.1016/j.cub.2011.01.011

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

Westheimer, G. (2005). Anisotropies in peripheral vernier acuity. *Spatial Vision*, *18*(2), 159–167.

Weymouth, F. W. (1958). Visual sensory units and the minimal angle of resolution. *American Journal of Ophthalmology*, *46*(1), 102–113.

Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *Journal of the Optical Society of America A*, *14*(9), 2057. https://doi.org/10.1364/JOSAA.14.002057

World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. (2013). *JAMA*, *310*(20), 2191. https://doi.org/10.1001/jama.2013.281053

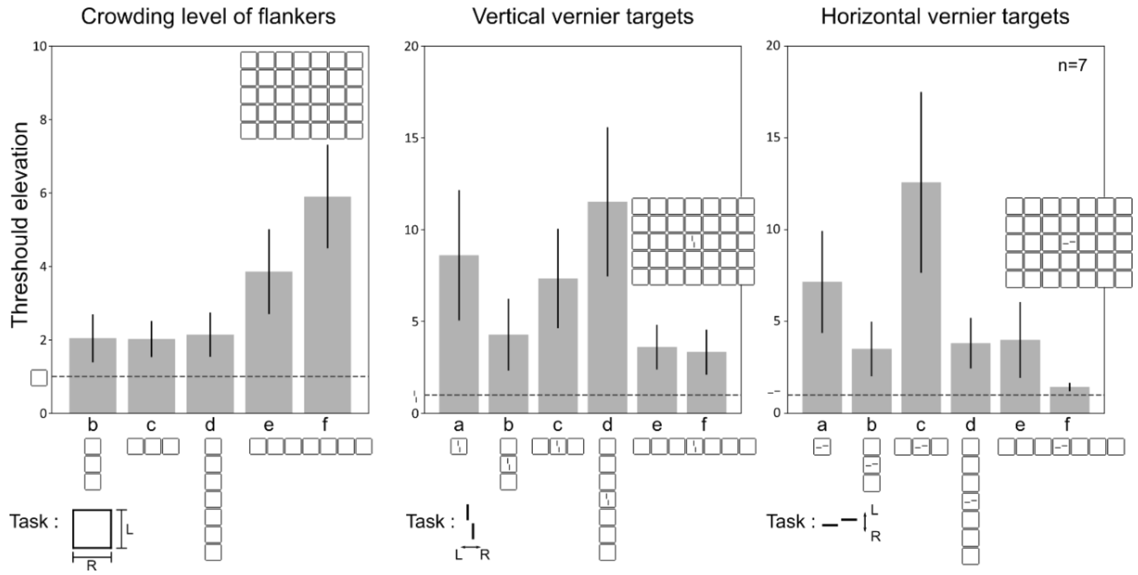
Xu, X., Collins, C. E., Khaytin, I., Kaas, J. H., & Casagrande, V. A. (2006). Unequal representation of cardinal vs. Oblique orientations in the middle temporal visual area. *Proceedings of the National Academy of Sciences*, *103*(46), 17490–17495.

Yeh, S.-L., He, S., & Cavanagh, P. (2012). Semantic priming from crowded words. *Psychological Science*, *23*(6), 608–616. https://doi.org/10.1177/0956797611434746

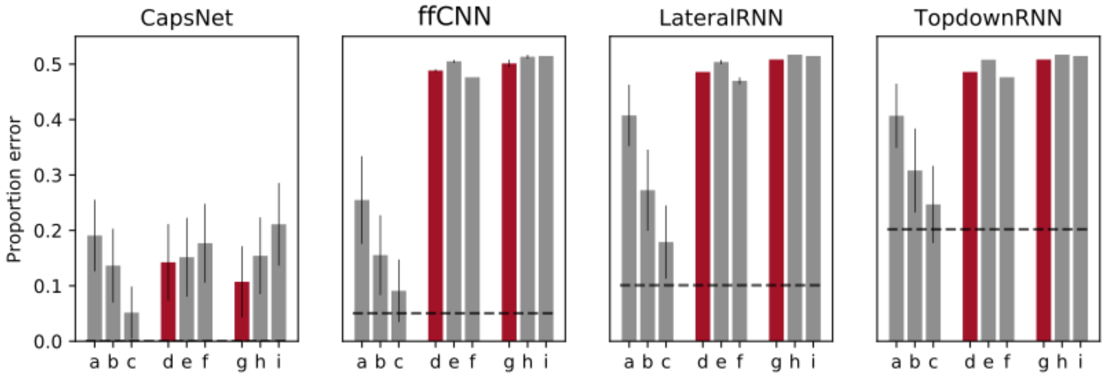
Yeotikar, N. S., Khuu, S. K., Asper, L. J., & Suttle, C. M. (2011). Configuration specificity of crowding in peripheral vision. *Vision Research*, *51*(11), 1239–1248. https://doi.org/10.1016/j.visres.2011.03.016

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. https://doi.org/10.1167/10.10.16

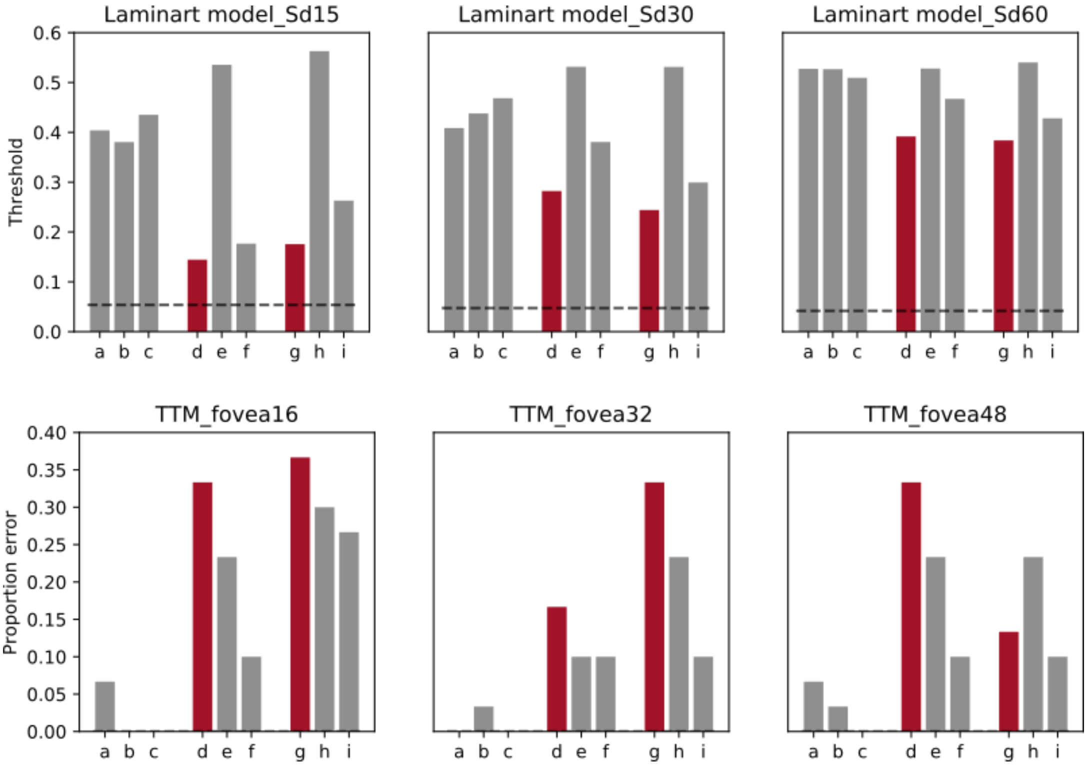
Supplementary Materials



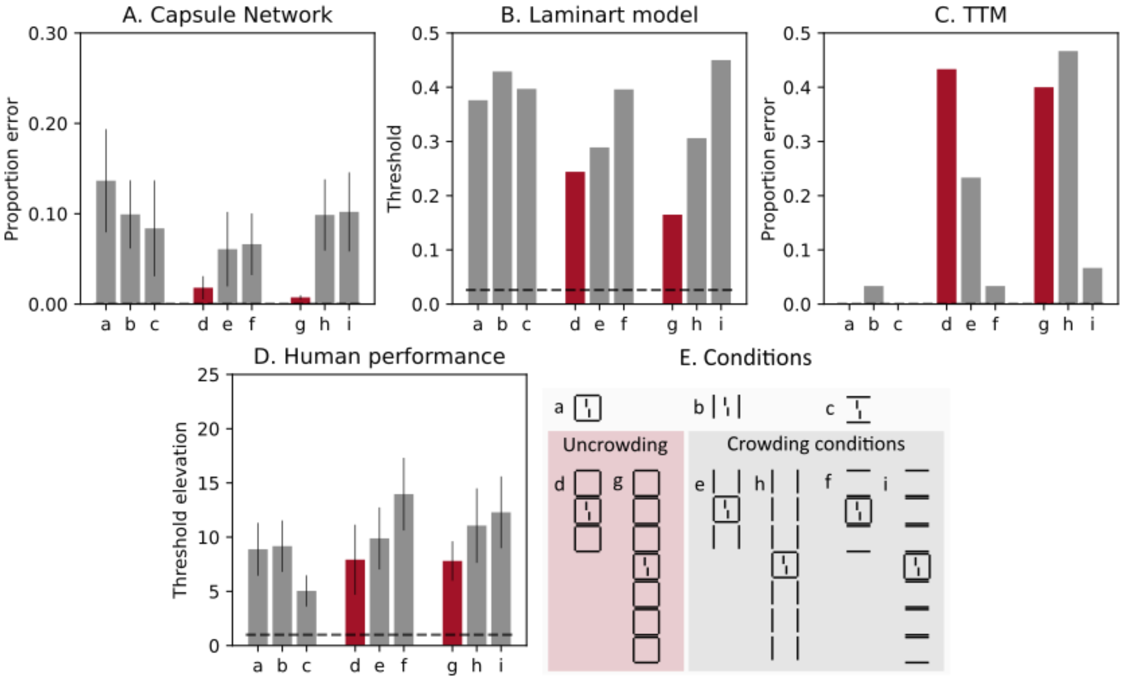
Supp. Figure 1. Experiment 3a results. The center square's ratio aspect discrimination task (left), the vertical vernier offset discrimination task (middle), and the horizontal vernier offset discrimination task (right). The results in the middle and the right panel were pooled in the pooled condition. In the left panel, the performance deteriorated (the target was more crowded) as the number of squares increased. The y-axis represents threshold elevation relative to the one square condition (left) and threshold elevation relative to the vernier only condition for the middle and right panel. Mean ± SEM.



Supp. Figure 2. Model performances of control models of CapsNets. Y-axis represents proportion error, thus, larger values indicate worse performances. X-axis labels the tested conditions as same as in Fig.6. As in Fig. 6, red bars represent conditions leading to uncrowding in humans (good performance), and gray bars represent crowding (poor performance). Gray dashed lines show the model performance for the vernier only condition. Three control models of CapsNets, ffCNN, lateral RNN, and topdown RNN, did not reproduce human nor CapsNets' performances, which is, good performances in uncrowding conditions (red bars) and bad performances in crowding conditions (grey bars). Note that the control networks had the same number of layers and neurons and trained with the same training procedure, and the only differences were in the connectivity between the neurons in the last two layers.



Supp. Figure 3. Model performances with different hyperparameters. Y-axis represents threshold or proportion error, thus, larger values indicate worse performances. X-axis labels the tested conditions as same as in Fig.6. As in Fig. 6, red bars represent conditions leading to uncrowding in humans (good performance), and gray bars represent crowding (poor performance). Gray dashed lines show the model performance for the vernier only condition. Free parameter for the Laminart model was extent of the selection signal window, and the fovea size for the TTM. In Fig. 6, we used the default parameters from the authors (TTM: fovea size=32 pixel (1 arcdeg), Laminart: sd=30 (2arcdeg)). Here, we reduced and increased values of the free parameters to confirm the model results were not merely artifacts from the hyperparameter settings. Changes in the parameters did not lead to obvious differences. Except that TTM with larger fovea size (fovea=48), the performance for 7 horizontal squares was better than that of 3 squares, but it was still worse than that of 1 square condition, thus, no uncrowding.



Supp. Figure 4. Model performances with vertically aligned flanker configurations. Y-axis represents threshold or proportion error, thus, larger values indicate worse performances. As in Fig. 6, red bars represent conditions leading to uncrowding in humans (good performance), and gray bars represent crowding (poor performance). Gray dashed lines show the model performance for the vernier only condition. The results were comparable to that of horizontally organized configurations (Fig. 6). The 2- stage models, Capsule networks and Laminart model, reproduced human performance, i.e., good performance in uncrowding conditions (red bars, conditions d and g), but strong crowding in crowding conditions (gray bars, conditions e and h). But the 1-stage-model, TTM, did not reproduce human performance, i.e., strong crowding in all the multiple flankers conditions.

Supplementary Tables *Parameter estimates of Linear Mixed Effects Models (LMMs)*

Table 1. Estimates from the linear mixed-effects model of pooled data. with the number of squares in the horizontal and the vertical dimensions as predictors (no interaction between the two predictors) and individual participants and target orientation as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	2.295	0.195	11.788
The num of columns of squares	-0.151	0.024	-6.269
The num of rows of squares	-0.170	0.021	-8.214

Table 2. Estimates from the linear mixed-effects model of subset of pooled data (3 and 7 flankers configurations). With the number of squares in the horizontal and the vertical dimensions and flanker orientations as predictors and individual participants and target orientation as random intercepts. The model explains 36.3% of the variance, and only 6.7% when not accounting for the random effects (r_m^2=0.067, r_c^2=0.363).

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	2.567	0.319	8.042
Flanker orientations	-0.969	0.306	-3.166
The num of columns of squares	-0.228	0.048	-4.714
The num of rows of squares	-0.019	0.048	-0.390

Table 3. Estimates from the linear mixed-effects model of 1 flanker conditions in Exp1. with the flanker configurations as predictors (holistic square used as a reference level) and individual participants and target orientation as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	1.915	0.232	8.256
Iso-target orientation flankers	-0.028	0.283	-0.099
Aniso-target orientation flankers	-1.199	0.283	-4.235

Table 4. Estimates from the linear mixed-effects model of 3 or 7 flanker conditions in Exp1. with the flanker configurations as predictors (holistic square used as a reference level) and individual participants and target orientation as random intercepts.

Fixed Effects	β estimate	β standard error	t-value

(Intercept)	1.262	0.248	5.096
Iso-target orientation flankers	0.696	0.138	5.060
Aniso-target orientation flankers	0.993	0.138	7.220

Table 5. Estimates from the linear mixed-effects model of vernier only conditions in Exp2. with the stimulus orientations as a predictor and individual participants and target orientation as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	4.361	0.148	29.485
Oblique orientation	0.994	0.209	4.753

Table 6. Estimates from the linear mixed-effects model of Exp2. with the number of squares as a predictor and individual participants and target orientation as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	1.858	0.155	11.969
The num of squares	-0.041	0.007	-5.991

Table 7. Estimates from the linear mixed-effects model of Exp3a. with the number of squares in the horizontal and the vertical dimensions as predictors (no interaction between the two predictors) and individual participants as a random intercept.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	-0.191	0.372	-0.513
The num of columns of squares	0.087	0.046	1.917
The num of rows of squares	0.162	0.039	4.130

Table 8. Estimates from the linear mixed-effects model of Exp3b. with the number of squares in the horizontal and the vertical dimensions as predictors (no interaction between the two predictors) and individual participants and cue existence as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	0.684	0.204	3.346
The num of columns of squares	-0.009	0.026	-0.341
The num of rows of squares	0.126	0.023	5.524

Table 9. Estimates from the linear mixed-effects model of 3,5, or 7 squares in Exp3b. with the number of squares and the flanker orientations as predictors (no interaction between the two predictors).

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	1.214	0.248	4.896
The num of squares	0.013	0.036	0.350
vertically aligned flankers	-0.524	0.119	-4.404

Model details

Capsule networks are deep learning systems in which layers of neurons communicate through a recurrent “routing by agreement” process (Sabour et al., 2017). The entire network is trained end to end through backpropagation. After training, each capsule is a group of neurons representing a certain feature. During routing by agreement, each capsule in the lower layer predicts the activity of each capsule in the next layer. When many capsules agree that a certain higher-level capsule should be highly active, the corresponding higher-level capsule is activated. For example, if a low-level capsule detects a rectangle, and another low-level capsule above detects a triangle, they agree that the higher-level object should be a house and, therefore, the corresponding high-level capsule is activated. Through this process, capsule networks can group and segment objects: the triangle and rectangle are grouped together and segmented into the house capsule. Doerig et al. (2020) showed that capsule networks can explain uncrowding based on these grouping and segmentation capabilities. Here, we tested if our results could be explained in this way, too. We trained and tested the capsule networks using exactly the same procedure as Doerig et al. (2020). We trained the networks to recognize verniers and group of squares, group of vertical bars, or group of horizontal bars. Hence, after the training, the networks could recognize these shapes, when presented in isolation from each other, but they had never seen vernier overlapped with other shapes. During testing, we quantified how well the Vernier offset could be decoded in the presence of different flanking configurations. We found that our human data was well explained by these capsule networks: increasing the number of squares improves the model performance (Fig. 6A. red bars), but this effect is absent with the vertical bars or horizontal bars (Fig. 6A. gray bars).

The Laminart model groups elements in the visual field by computing illusory contours between collinear edges. Then, local selection signals trigger a segmentation process that spreads along connected (illusory or real) contours (Francis et al., 2017). Therefore, if hit by a selection signal, all elements linked by illusory contours are segmented to a distinct neural population and processed separately. Doerig, Bornet and colleagues (2019) had shown that local selection signals trigger the segmentation better by the number of squares increases, which led to the stronger uncrowding. Here, we used the same model and the same evaluation method to derive the model performance for each stimulus. In the single flanker conditions (Fig. 6B a, b, c), the vernier target and the flanker are close to each other and hence, in most of the cases, the same selection signal will hit both of them. In consequence, both flanker and target are processed as a single perceptual group and interference happens between them. For this reason, crowding is strong. In the multiple squares conditions (Fig 6B d, g), the squares are connected to each other by illusory contours. These squares span a larger region than in the single flanker condition. Hence, the selection signals can easily hit the group of flankers without hitting the vernier target. In consequence, the flankers are in most of the cases segregated from the target. Hence, no interference happens between the flankers and the target, reducing crowding strength (uncrowding, good performance). In the multiple horizontal lines conditions (Fig. 6B f, i), the flankers still induce illusory contours and connect with the central square. These conditions are thus equivalent to the multiple squares conditions and lead to uncrowding. However, in the multiple vertical lines conditions (Fig 6B e, h), the flankers do not form illusory contours that reach the central square. These conditions are thus equivalent to the single square condition and lead to strong crowding (bad performance).

The TTM, unlike other two-stages models, describes visual perception as an image statistics rather than by the pre-segmented texture elements or objects (Portilla & Simoncelli, 2000; Rosenholtz, 2014). Therefore, the model pools information over each local pooling region, which grows linearly with eccentricity, and then tiles pooled information over the whole visual field. Here, we evaluated the partial square conditions in addition to the square configurations tested by Doerig, Bornet and colleagues (2019). In addition to Doerig and colleagues (2019), the TTM showed no uncrowding in neither partial nor full squares, of the configurations. Instead, the performance showed an interesting pattern. Model performance with full squares was worse than with partial squares (i.e., Fig. 6C g vs. i), contrary to the human performance (Fig. 6D g vs. i), but in consistent with Rosenholtz and colleagues (2019). They speculated that less crowding in ‘Uncrowding’ stimuli is due to the decision bias induced by target dissimilar flankers, but not perception per se. For instance, in a configuration like Fig. 5e&i, target position uncertainty is reduced thanks to the horizontal lines creating a different configuration from the target and the center square. Thus, such a difference works as a ‘cue’ to bias the decision towards the target position. However, unlike the TTM, human performance was largely deteriorated in ‘cued’ (partial square) configurations than in ‘un-cued’ (full square) configurations (Fig. 6D g vs. h,i).

3.4 Basic Gestalt principles cannot explain (un)crowding

Full citation: **Choung, O. H.**, Rashal, E., Kunchulia, M., & Herzog, M. H. (*in prep*). Basic gestalt principles cannot explain Uncrowding.

Summary:

Gestalt principles are considered as the fundamentals of perceptual organization. In this study, we asked if basic Gestalt principles can explain (un)crowding. Here, we used shape similarity as the basis for all the other grouping principles. We generated 40 configurations that follow ten Gestalt principles to examine whether basic Gestalt principles can explain (un)crowding. We expected that certain principles might contribute to the grouping stronger than the other principles. For example, flanker configurations with 2 symmetry axes may always show good performances, whereas configurations with irregular good continuation do not. In contrast to our expectations, performance was hardly explained by Gestalt principles. The performances related to the same Gestalt principle were not consistent. For example, some configurations with 2 symmetry axes showed strong uncrowding, whereas others showed strong crowding.

Then, we asked whether grouping or perceptual organization matter at all in uncrowding. We measured five subjective measures for grouping and segmentation to examine the correlations with (un)crowding performance. Grouping and segmentation measures were significantly correlated with (un)crowding performance even after Bonferroni correction. However, subjective grouping and segmentation ratings could not be explained by basic Gestalt principles. Similar to (un)crowding, no single rule had high subjective rates or low subjective rates in general.

Altogether, these results indicate that subjective ratings of grouping are highly correlated with (un)crowding performance. However, grouping processes could not be explained by any specific basic Gestalt principle.

Manuscript in preparation:

Basic gestalt principles cannot explain (Un)crowding

Oh-Hyeon Choung[a], Einat Rashal[a,b], Marina Kunchulia[c,d], Michael H. Herzog[a]

[a] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, http://lpsy.epfl.ch, ohhyeon.choung@gmail.com, michael.herzog@epfl.ch

[b] Department of Experimental Psychology, University of Ghent, Belgium, einatrashal@gmail.com

[c] Vision Research Laboratory, Beritashvili Centre of Experimental Biomedicine, Tbilisi, Georgia

[d] Institute of Cognitive Neurosciences, Agricultural University of Georgia, Tbilisi, Georgia

Abstract

Uncrowding cannot happen without grouping and segmentation. But, then, what makes a group? Gestalt principles, such as similarity, collinearity, etc., are the classic type of explanation. Here we used multiple Gestalt principles to test. We did not find a clear link between Gestalt principles and (un)crowding. However, the subjective ratings and the number of the directly connected objects may explain.

Introduction

In crowding, target perception strongly deteriorate when presented within clutter. Crowding is a major challenge for vision since elements are never presented alone. Traditionally, crowding is thought to be stronger when target and flanker features are similar, e.g., sharing the same color, motion, contrast, etc. Crowding is thought to happen only when flankers fall within a restricted area, around the target, which is often said to be half of the target's eccentricity (Bouma's window; Bouma, 1970; Bouma, 1973; Levi, 2008; Strasburger et al., 1991; Weymouth, 1958). Accordingly, crowding is often explained by feature pooling or averaging (e.g., Dakin et al., 2010; Greenwood et al., 2009, 2017; Parkes et al., 2001; Pelli, 2008; Rosenholtz et al., 2012; Solomon et al., 2004).

However, feature similarity is not determinant for crowding. For example, despite, indeed, green flankers do little harm to a red line target, red-green alternating flankers can exert strong crowding (Manassi et al., 2012; Sayim et al., 2008). Crowding is also not restricted to a fixed region but depends on the exact stimulus configuration. Flankers outside Bouma's window can deteriorate perception (Manassi et al., 2012; Rosen & Pelli, 2015; Saarela et al., 2010; Vickery et al., 2009). Interestingly, adding flankers within and outside Bouma's window can improve performance, that is, a release of crowding Chicherov et al., 2014; Chicherov & Herzog, 2015; Doerig et al., 2019; Herzog et al., 2015, 2016; Herzog & Manassi, 2015; Malania et al., 2007; Manassi et al., 2012, 2013, 2015, 2016; Saarela et al., 2009; Sayim et al., 2008, 2010, 2011; Choung et al., 2021). Two bars are presented below another with a small gap and an offset between the bars (Vernier). It is easy to discriminate the offset direction (left or right). Vernier offset discrimination drops drastically when the vernier is presented within a square (crowding). However, adding more squares improves performance almost to the unflanked, Vernier alone, condition's level (uncrowding). The Vernier information is recovered likely because, the target and the number of squares are in different perceptual groups, which is not the case when only one square is presented. Manassi and colleagues (2013) have shown that the target is easily perceived when the central square is crowded by the other squares (also see Choung et al., 2021). The target recovery seems largely correlated with flanker suppression (also see flanker suppression by awareness (Wallis et al., 2019) and masking (Chakravarthi & Cavanagh, 2009)).

As shown by Doerig and colleagues (2020), one-stage, feedforward models cannot explain uncrowding since target information is irretrievably lost in low-level processing. This holds true for local pooling models (e.g., Dakin et al., 2010; Greenwood et al., 2009, 2017; Parkes et al., 2001; Pelli, 2008; Rosenholtz et al., 2012; Solomon et al., 2004), and also including the models that can account for the global configurations, such as high dimensional feature pooling model (HD pooling; Bornet et al., 2021; Choung et al., 2021; Rosenholtz et al., 2019), deep networks (DNN; Doerig, Bornet, et al., 2020). Instead, it seems that recurrent processing that segregates different perceptual groups is needed, such as used in Capsule networks and the Laminart model (Doerig et al., 2019; Doerig, Schmittwilken, et al., 2020). However, these models cannot explain the complex stimuli. It remains unclear which aspects of the configuration induce grouping.In other words, the specific mechanism of grouping needs to be studied.

Here, we addressed this question by asking whether classic Gestalt principles can explain the central grouping aspects in crowding.

Manassi and colleagues (2012) have shown that crowding is correlated with the participants' subjective ratings about how strongly the vernier target stand-out from the flanker configuration, but this was only an indirect measure of grouping and segmentation. Hence, we assessed indirect and direct subjective measures to see to what extent they correlate with crowding.

Gestalt principles, which in German means 'formation', have been studied over centuries and is considered the fundamental of perceptual grouping (von Ehrenfels, 1890; Wertheimer, 1912, 1922, 1923; Köhler, 1920; Koffka, 1935; Metzger, 1936, 2006; reviews: Todorović, 2007; Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al., 2012). This is often summarized by: "The whole is more than the sum of its parts" (Aristotle, BC 384-322). In other words, perception of the overall configuration cannot be explained by a mere combination of individual components. Here are suggested Gestalt principles for grouping: symmetry, proximity, similarity, common fate, good continuation, closure, symmetry, parallelism, synchrony, common region, element, uniform connectedness. A single or the combination of the listed principles may define the uncrowding performance.

There are evidences of grouping cues that lead to segmentation. For example, Kovács and Julesz (1993) showed that collinearity and closure make gabor patches within the closed circle segments out from the other gabor patches. The other evidences include proximity (Claessens & Wagemans, 2005, 2008), collinearity (Hadad et al., 2010), similarity (Casco et al., 2009), etc.

Here, we used shape similarity as the basis for all the other grouping principles. We generated 40 configurations that follow ten Gestalt principles. Moreover, we measured five subjective measures for grouping and segmentation.

Materials and Methods

Participants

Thirty-one participants took part in the experiment. Eleven out of the 31 participants were excluded right after the calibration session, because they did not show strong crowding in the one square condition, which is a prerequisite for a release of crowding, to avoid the ceiling effect (see *Calibration session*). Hence, we retained the data of twenty participants (mean age: 21.6±1.6, 10 females, all right-handed, 7 with left eye dominance). All participants had normal or corrected to normal visual acuity in the Freiburg Visual Acuity Test as indicated by a binocular score greater than 1.0 (Bach, 1996). Observers gave written consent before the experiments. All experiments were conducted in accordance with the Declaration of Helsinki except for preregistration (World Medical Organization, 2013) and were approved by the local ethics committee (Commission éthique du Canton du Vaud).

Apparatus

Stimuli were displayed on a gamma-calibrated 24 inch ASUS VG248QE LCD monitor (1920x1080 px, 120 Hz). The room was dimly illuminated (~ 0.5 lux). Viewing distance was 75cm and the participant's chin and forehead were positioned on a chin-rest. Responses were collected using wireless hand-held push buttons. In the Vernier discrimination task, when no response was registered within 3 seconds, the trial was repeated randomly within the same block. A feedback tone was given for incorrect responses (high tone, 600 Hz) and omissions (low tone, 300 Hz).

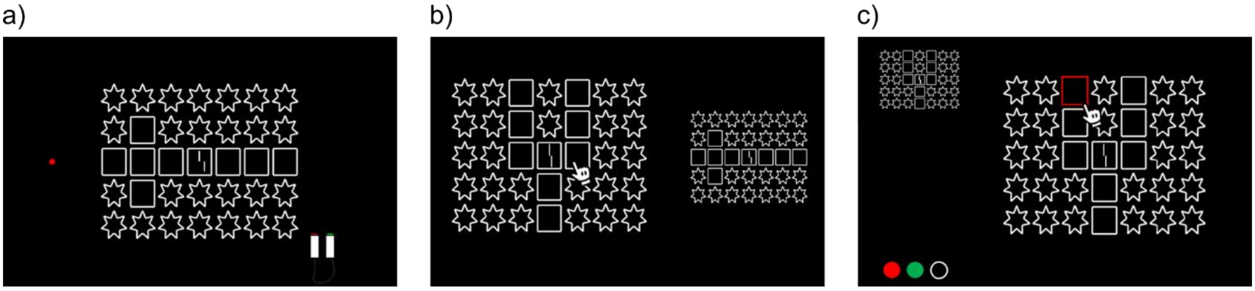


Figure 1. a) VCrowd task: Vernier offset discrimination task. The task was to discriminate whether the lower Vernier bar is offset to the left or right compared to the upper bar. b) VRank task: Vernier stand-out ranking task. Two stimuli were presented on the screen side-by-side with the same size.

The task was to choose from which flanker configuration the Vernier target stands out more strongly. All possible comparosons, i.e., 20*39 pairs of configurations were tested. c) Rate task: Participants were asked to rate how much the Vernier stands out, how much the center group stands out from the other elements, and how strongly the center group elements group with each other.

General procedures

Three tasks (Fig. 1) were carried out with forty flanker configurations (Fig. 2). The three tasks were a vernier discrimination task (VCrowd), a vernier stand-out ranking task (VRank), and a rating task (Rate). VRank and Rate tasks were tested twice. The experiment was conducted on 5 days within 2 weeks (day 1-3: calibration and VCrowd, day 4: VRank twice and Rate, day 5: Rate). All the participants went through a calibration session to adjust flanker size individually.

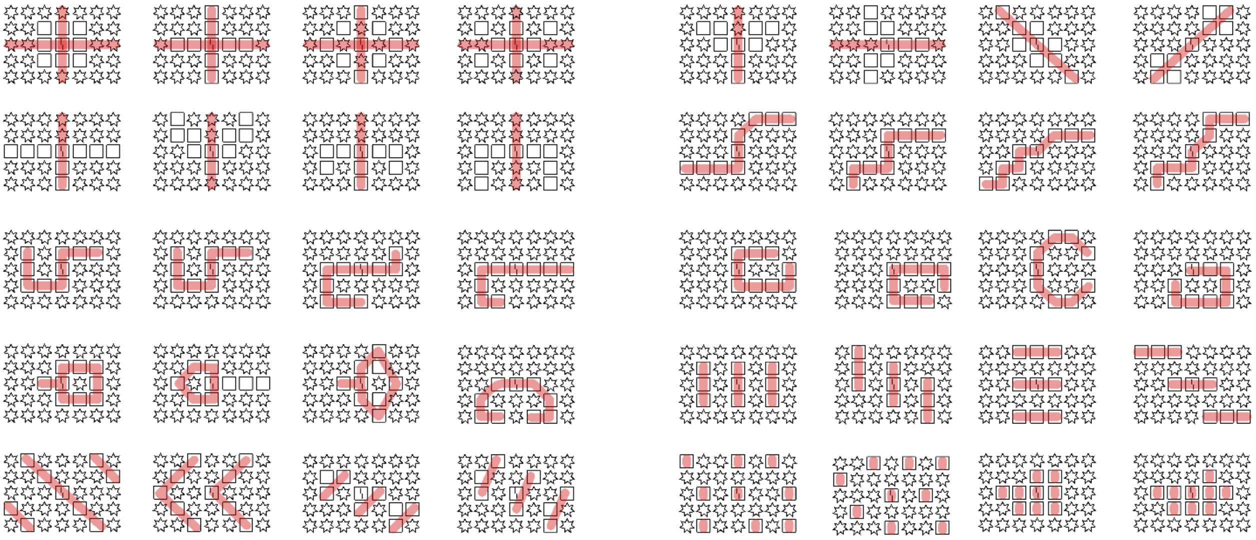


Figure 2. Flanker configurations. Red lines indicate the tested Gestalt principle; these lines are for illustration only and were not presented during the experiment. There were four configurations for each Gestalt principle: row 1 left- symmetry with 2 axis (Symm2); row 1 right – symmetry with 1 axis (Symm1); row2 left– symmetry with 1 axis (Symm1); row2 right – collinear good continuation (CollCont); row 3 left-irregular good continuation (Irre-Cont); row3 right-closure (Close); row4 closure with symmetry (SymmClose); row4 right-repetition (Repeat); row5 left-repetition diagonal (RepeatDia); row5 right-random without (first 2: RandWithout) and with (last 2: RandWith) group. Note that RandWith configurations could be considered as grouping by proximity. Most of the configurations were composed of 9 squares and 26 stars (* three exceptional configurations which had 10 squares and 25 stars). Therefore, low-level features, such as pixel values, were semi-identical across the configurations.

Stimuli

Stimuli were white (100 cd/m^2) and were presented on a black background with luminance below 0.3 cd/m^2. Participants were asked to fixate on a red fixation dot (diameter = 8 arcmin, 20 cd/m^2). Each stimulus was composed of a Vernier target and the differently configured flankers. The Vernier target was composed of two 40 arcmin long 1.8 arcmin wide bars vertically presented one below the other, and the gap between two bars was 4 arcmin. Left/right offsets were balanced within a block. The Vernier target was surrounded by 35 flanker elements, which were composed of 9 squares and 26 stars (except 3 configurations). Squares and stars were positioned in 5 rows and 7 columns as in Fig. 2, and there were 40 different configurations. Each flanker configuration followed one of four Gestalt principles; symmetry (in 2 axes or 1 axis), good continuation (regular or irregular), closure (regular or irregular), repetition (regular or irregular), or without following any rules (with and without group). Note that the center flanker was always a square, and the Vernier target was always located in the center. Except for the VCrowd task, each square was composed of four 120 arcmin long lines and each star was composed of seven 48 arcmin long lines. The center-to-center distance between flankers was 150 arcmin. For the VCrowd task, the square and star sizes and the gap between flankers were individually adjusted in a calibration session (details in *Calibration session*). The side length of the squares was 84 – 114 arcmin, and the gap between squares was 21-28.5 arcmin depending on observers.

Each configuration was presented at the center of the screen, and the fixation dot was presented at an eccentricity of 9 degrees to the left. Hence, stimuli were presented at 9 degrees in the visual periphery. The chin-rest was always placed 75cm from the fixation dot. Psychophysics Toolbox was used to present the stimuli (Brainard, 1997; Kleiner et al., 2007; Pelli & Vision, 1997). To avoid viusal aftereffects, a small spatial jitter was applied to the entire stimulus within a 3 pixels range from trial to trial.

Procedures

Calibration session. Before starting the main experiment, to avoid floor and ceiling effects, each participant went through a calibration session. The calibration session was composed of two conditions. First, 1 or 2 blocks with the Vernier alone condition (160 trial per block) were tested to familiarize with the Vernier task (participants with threshold larger than 200 arcsec went through the 2nd block). Second, up to 7 blocks with a vernier surrounded by one square (80 trial/block) were tested to find the spatial parameters so that performance with one surrounding square was at least 6 times larger than the performance in the Vernier alone condition (mean thresholds for the Vernier alone condition: 142.30 ± 45.48, one square condition: 935.84 ± 188.53). We reduced the flanker size and the flanker-to-flanker distance gradually, until the threshold of the one square condition reached at least 6 times the threshold of the Vernier alone condition. We excluded participants whose threshold were still below the criterion even after reducing the square size to 70%. In total, 11 of 31 participants were excluded.

VCrowd task *(the vernier discrimination task; Fig.1a)*, the stimulus (Vernier + flankers) was presented for 150ms in the center of the screen, and participants were asked to discriminate whether the lower bar was on the left or right compared to the upper bar, by pressing the left or right button. Each configuration was tested in a block of 100 trials. To reduce target-location uncertainty, only the target was presented alone for 150ms at the beginning of each block. We used the PEST (Parameter Estimation by Sequential Testing) stair-case procedure (Taylor & Creelman, 1967) to determine testing levels (offsets). PEST procedure changes test levels depending on the recent history step-wise. Therefore, the test levels are only changed when the hit rate lies above or below a threshold criterion of 75%. The step-sizes are optimized to the hit rate coverage to 75%. The procedure ended after 100 trials, and the thresholds (*Thresh*) was derived from post-hoc psychometric function fittings to the data (details in *Data analysis*).

VRank task *(the Vernier stand-out ranking task; Fig.1b)*, two same-sized flanker configurations were presented simultaneously side-by-side, and participants were asked to choose in which flanker configuration the Vernier target stands out more strongly, i.e., a "win" (Fig. 1). Overall, 718 (20*39) pairs of configurations were tested twice. The responses from two blocks (2*718) were combined and used for evaluating the rank (Rank). For each configuration we counted the "wins". When two or more configurations had the same number of 'wins', i.e., a "draw", we counted the 'wins' among the "draw" configurations and evaluated the rank (Rank). That is, the more 'wins' the configuration had, the higher the rank it achieved.Thus, for each participant,we have a rank order of the configurations ranking from1 to 40. In addition to the individual Rank order per participant, a global rank (GlobRank) was obtained by using a similar process, by pooling the responses from twenty participants.

Rate task *(the rating tasks; Fig.1c)*, four questions were asked. First, as in the VCrowd task, the stimulus was presented for 150 ms, and participants were asked to rate how much the vernier target stands out from the flanker configuration on a scale from 1 to 5 (*VStandRate*). Second, the stimulus was presented with unlimited time, and the participants were asked to assigm each flanker element to different sub-groups, based on their perception. Then, they were asked to rate two questions on a scale from 1 to 5; 1) how much does each sub-group stands out from the other groups (*GStandRate*), and 2) how strongly do the elements in each group grouped together (GGroupRate)?

Hence, we determined five measures: crowding threshold (Thresh; from the VCrowd task), vernier stand-out ranking (Rank; from the VRank task), vernier stand-out rating (VStandRate; from the Rate task), group stand-out rating (GStandRate; from the Rate task), and grouping strength (GGroupRate from Rate task).

Data analysis

We fitted a cumulative Gaussian function to the data and determined the vernier offset threshold (Thresh) for which 75% correct responses were reached. High thresholds indicate inferior performance, low thresholds indicate good performance. The Psignifit 2.5 toolbox (Fründ et al., 2011) was used for psychometric function fitting. We computed threshold elevation for each condition and each observer, i.e., we divided the threshold in each condition by the threshold in the Vernier alone condition. Data were log transformed to bring the data closer to normality. No obvious violation was detected by visual inspection.

Using R (R Core Team, 2019) and *lme4* package (Bates et al., 2015), we computed linear mixed-effects models (LMM) to account for random variations due to individual differences. The fixed and random effects specified for each experiment. The model significance (*p*-value) was obtained through likelihood ratio tests (χ^2*)* by comparing nested models. For each fitted model, using *MuMIn* package (Barton, 2020), we computed the effect size (r^2), i.e. the explained variance, when including (conditional r_c^2) and excluding (marginal r_m^2) the random effects (Johnson, 2014; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013). Posthoc multiple comparisons of means were computed with *multcomp* package (Hothorn et al., 2008).

Intra-rater reliability for Rate task was carried out by using ordinal alpha (Zumbo et al., 2007) to account for ordinality of the measures (VStandRate, GStandRate, and GGroupRate). psych (Revelle, 2021) package was used.

Correlations between the measures were computed using Spearman rank correlation (Spearman, 1904), as four measures among five were in ordinal scale. Moreover, to account for the individual variances, the significance of the correlations was obtained through the radomization test (details in Supp. Method; Bakdash & Marusich, 2017; Mohr & Marcon, 2005).

Model simulations

Two model bases were tested. The first model tests whether the number of direct neighboring squares and discounted by the number of star flankers depending on their distances can explain uncrowding. The second model tests whether pixels within the crowding window can determine crowding.

The models serve as a test for simple low vs. high level features.

1) Directly connected squares' and flankers' euclidean distance

Model performance was predicted by the inverse of the averaged Euclidean distance from the center square to each of the directly connected square. Therefore, the closer the connected square is, the better the performance is. The performance s_i of configuration (condition) *i* was computed as follow.

$$s_i = M_i \quad \text{(1: Num\_sq)}$$

$$s_i = M_i \times id_i,\ where\ \ id_i = \sum_j^{M_i} \frac{1}{\sqrt{r_{ij}^2 + c_{ij}^2}} \qquad \text{(2: Num\_sq\_group)}$$

$$\text{Or} \quad s_i = -T_i \times f_i,\ where\ \ f_i = \sum_j^{T_i} \frac{1}{\sqrt{r_{ij}^2 + c_{ij}^2}} \qquad \text{(3: Num\_sq\_flankers)}$$

Where $M_i \in [1, 10]$ denotes the number of directly connected squares to the center square for configuration (condition) i, and r_{ij} and c_{ij} encodes row and column numbers from the center; $T_i \in [15, 24]$ encodes the number of flankers except the directly connected squares ($35 - M_i$) for each configuration (i); the center square is annotated as $(0,0)$, $r_{ij} \in [-2, 2]$ and $c_{ij} \in [-3, 3]$. We tested three variances of the number of directly connected squares based model. 1) Num_sq: merely the number of squares as the predictor; 2) Num_sq_group: the number of squares discounted by inverse of sum of distance; 3) Num_sq_flankers: minus the number of non-connected square flankers discounted by inverse of sum of distances.

2) Pixel-wise euclidean distance

As a control, we also tested whether flankers' pixel values all over the configuration or within the fixed crowding window (1/2 of eccentricity) can predict the performance, which is the traditional view of the crowding (pooling; see REF). Similar to Eq. 1, the performance s_i was computed as follow.

$$s_i = M_i \times id_i,\ where\ id_i = \sum_j^{M_i} \frac{1}{\sqrt{x_{ij}^2 + y_{ij}^2}} \qquad \text{(4: Pix\_n\_all)}$$

$$of\ which \quad \frac{x_i^2}{a^2} + \frac{y_i^2}{b^2} = 1, \quad a = \frac{1}{2}ecc, b = \frac{1}{4}ecc \qquad \text{(5: Pix\_n\_bouma)}$$

M_i encodes the number of flanker pixels in configuration (condition) i, and x_i and y_i are the pixel position from the screen center $(0,0)$. Pixel-wise distance measure was computed within the Bouma's window, which is an ellipse with one focal is ½ of eccentricity ($a = 4.5deg$) and the other focal point is ¼ eccentricity ($b = 2.25deg$). Three variences of the model were tested; 1) Pix_num: merely the number of pixels; 2) Pix_n_all: the number of pixels discounted by inverse of the sum of pixel distances; 3) Pix_n_Bouma: same as Pix_n_all but only pixcels within the Bouma's window.

3) Model significance test

We analyzed the predictability of the models using two methods. First, we used LMMs which had each of the model estimates as fixed effects. For each LMM, the fixed effect was model estimates for each configuration, and each subject was considered as random intercepts.

Next, we used a leave one out cross validation (LOOCV) method to determine the explained variance of participants' performance. We linearly fitted the model estimates to the crowding performance of 19 participants behavioral data. Then, we obtained the r squared value (explained variance) by using the last participants' data (data points are not included in the linear regression). We repeated the compuation 20 times (for each participant), then averaged the r squared values from 20 iterations to get the final explained variance of each model.

Results

Intra-rater reliability

We computed ordinal alpha (Gadermann et al., 2012; Zumbo et al., 2007) to test the reliability of repeated responses for the three measures (VStandRate, GStandRate, GGroupRate) for the Rate Task. Most of the measures showed substantial reliability for all configurations with an alpha larger than 0.7 (Cohen, 1988; McHugh, 2012) VStandRate: $\alpha \in [0.730, 0.992]$; GStandRate: $\alpha \in [0.708, 1]$; GGroupRate: $\alpha \in [0.595, 1]$), except for two configurations in GGroupRate. For this reason, we used the averaged rating values in the subseuqent analyses.

Gestalt principles cannot explain (un)crowding

Here, we tested to what extent perceptual grouping can be explained by the Gestalt principles, and whether certain principles contribute stronger than others. For example, flanker configurations with 2 symmetry axes may show better performance than configurations with irregular good continuation (Fig. 3, red bars vs. orange bars), because symmetricity makes a more robust perceptual group. We tested forty configurations, which followed ten different Gestalt principles.

In contrast to our expectations, performance was hardly explained by Gestalt principles. Figure 3 shows performance for each configuration. Each color represents one of the ten Gestalt principles. Importantly, the performances related to the same Gestalt principle were not consistent. For example, some configurations with two symmetry axes showed uncrowding (red bars lower than the grey dotted line), whereas the other two showed crowding (red bar higher than the grey dotted line). We used a linear mixed effect model (LMM) with the fixed effect of Gestalt principles and random intercepts of configurations and participants. Although the fixed effect was significant (likelihood ratio test between models including and excluding the fixed effect: $\chi^2(9)=29.519$, $p<0.001$), no clear hierarchy among the Gestalt principles was found. Post-hoc Tukey's HSD comparison showed further evidence for that no single rule dominates performance (details in Supp. Table 1).

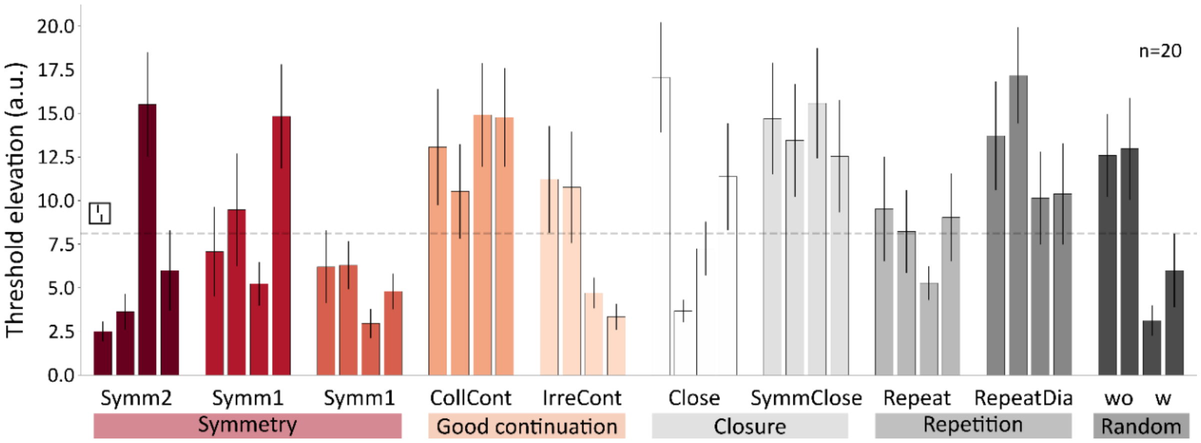


Figure 3. Performance for each configuration. Each color represents a different Gestalt principles. Configurations are presented in the same order as in Fig. 2.The y-axis shows threshold elevation compared to the Vernier only condition. Mean ± SEM.

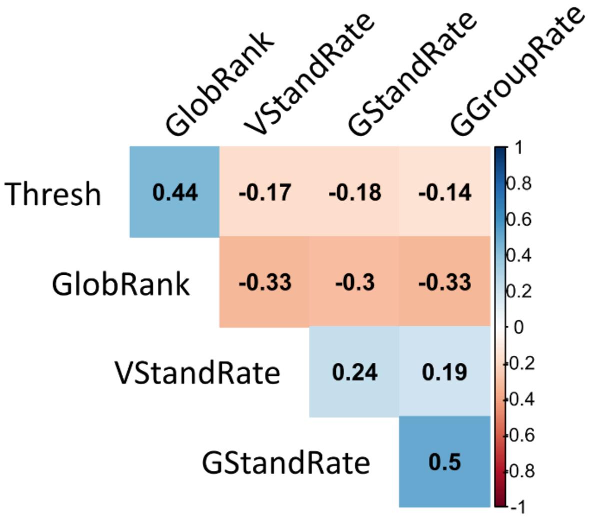
Subjective grouping and segmentation measures are correlated with crowding level but not with a specific principle

Gestalt principles could not fully explain the performance in the VCrowd task. Two reasons come to mind: 1) Gestalt principles are not a major driving source of flankers' grouping, and 2) (un)crowding is not mediated by grouping and segmentation.

First, we used LMMs to test if Gestalt principles can be a predictor for the grouping and segmentation measures (Rank, VStandRate, GStandRate, GGroupRate). An LMM with a fixed effect of Gestalt principles and random intercepts of configurations and participants was examined for each measure. All the models showed a significant fixed effect (Rank: $\chi^2(9)= 26.558$, $p_{Bonf}<0.01$; VStandRate: $\chi^2(9)= 28.704$, $p_{Bonf}<0.01$; GStandRate: $\chi^2(9)= 45.347$, $p_{Bonf}<0.001$; GGroupRate: $\chi^2(9)= 51.338$, $p_{Bonf}<0.001$; detailed estimates in Supp. Table 2). However, similar to (un)crowding, no single rule had high rates or low rates in general, except that configurations following RandWithout (configurations does not follow any Gestalt principle and does not have group; Fig. 2 row5 right last 2) had significantly worse ratings than the other rules (post-hoc Tukey's HSD test; details in Supp. Table 3). Our results suggest that Gestalt principles may contribute to grouping and segmentation, but the contribution is not exclusive. In other words, there are more factors than simply Gestalt principles.

Next, we tested correlations between the performance measure (Thresh) and the grouping and segmentation measures. As expected, all the measures had significant correlations even after the Bonferroni correction (Thresh-GlobRank: r_{mean}=0.44, p_{Bonf}<0.001, $95\%CI_{percent}$=[0.38, 0.49]; Thresh-VStandRate: r_{mean} =-0.17, p_{Bonf}<0.01, $95\%CI_{percent}$=[-0.23, -0.10]; Thresh-GStandRate: r_{mean} =-0.18, p_{Bonf}<0.01, $95\%CI_{percent}$=[-0.24, -0.11]; Thresh-GGroupRate: r_{mean}=-0.14, p_{Bonf}<0.01, $95\%CI_{percent}$=[-0.21, -0.07]; GlobRank-VStandRate: r_{mean}=-0.33, p_{Bonf}<0.001, $95\%CI_{percent}$=[-0.38, -0.35]; GlobRank-GStandRate: r_{mean}=-0.3, p_{Bonf}<0.001, $95\%CI_{percent}$=[-0.36, -0.23]; GlobRank-GGroupRate: r_{mean}=-0.33, p_{Bonf}<0.001, $95\%CI_{percent}$=[-0.39, -0.26]; VStandRate-GStandRate: r_{mean}=0.24, p_{Bonf}<0.001, $95\%CI_{percent}$=[0.17, 0.30]; VStandRate-GGroupRate: r_{mean}=0.19, p_{Bonf}<0.001, $95\%CI_{percent}$=[0.13, 0.25]; GStandRate-GGroupRate: r_{mean}=0.5, p_{Bonf}<0.001, $95\%CI_{percent}$=[0.44, 0.55]). We computed Spearman's Rank correlation to account for the ordinal scales, significance was obtained by randomization tests (details in Supp. Methods). Figure 4 shows the average Spearman r coefficients. The full results for each configuration and the distributions of the randomization test are presented in Supp. Fig. 1. Interestingly, there were stronger correlations between each step. For instance, the correlation between (un)crowding (Thresh) and Vernier stand-out (GlobRank) measures was high; two Vernier stand-out measures had a strong correlation (GlobRank-VStandRate).

Altogether, these results indicate that subjective ratings of grouping and segmentation are indeed highly correlated with the (un)crowding performance. However, grouping processes could not be explained by any specific Gestalt principle.



	GlobRank	VStandRate	GStandRate	GGroupRate
Thresh	0.44	-0.17	-0.18	-0.14
GlobRank		-0.33	-0.3	-0.33
VStandRate			0.24	0.19
GStandRate				0.5

Figure 4. Correlations among measures. Color code represents the mean Spearman's rank coefficient. All the correlations were significant after the Bonferroni correction.

The number of repeated squares rather than pixel values may predict (un)crowding

There was a strong correlation between crowding and the subjective grouping and segmentation measure. Thus, what explains grouping then?

Francis and colleagues (Francis et al., 2017) suggested that the Laminart model can predict crowding and uncrowding by collinearity grouping, similar to 'association field', the aligned (or collinear) edges are connected to make the perceptual group. Choung and colleagues (2021) showed that such a 'association field' grouping is unlikely in humans. For example, human participants uncrowd the flankers when 7 horizontally aligned squares were presented together with the target. However, when only horizontal lines of the flanker squares were presented, instead of the complete square, human performance was poor, which was not the case for the Laminart model (See Fig. 6 of Choung et al., 2021).

Therefore, global processing such as the good Gestalt of squareness is needed; when the number of complete squares were aligned and presented, flankers are grouped together which leads target to be segmented out from the flankers. Similarly, Sayim and colleagues (2019) suggested that when an object is repeatedly presented in the same frame, some of the objects may be ignored. Here we suggest such an ignorance (or *redundancy masking*) of objects can be a major source of grouping the squares, when the squares are positioned close to each other, i.e., the center square is ignored (or masked) by repeatedley presented other squares.

We tested whether the number of directly connected squares (eq. 1-3) can predict the human performance. As controls, we also computed the low-level pixel values to examine whether simple pixel values (eq. 4 & 5) can explain the perforamance at all.

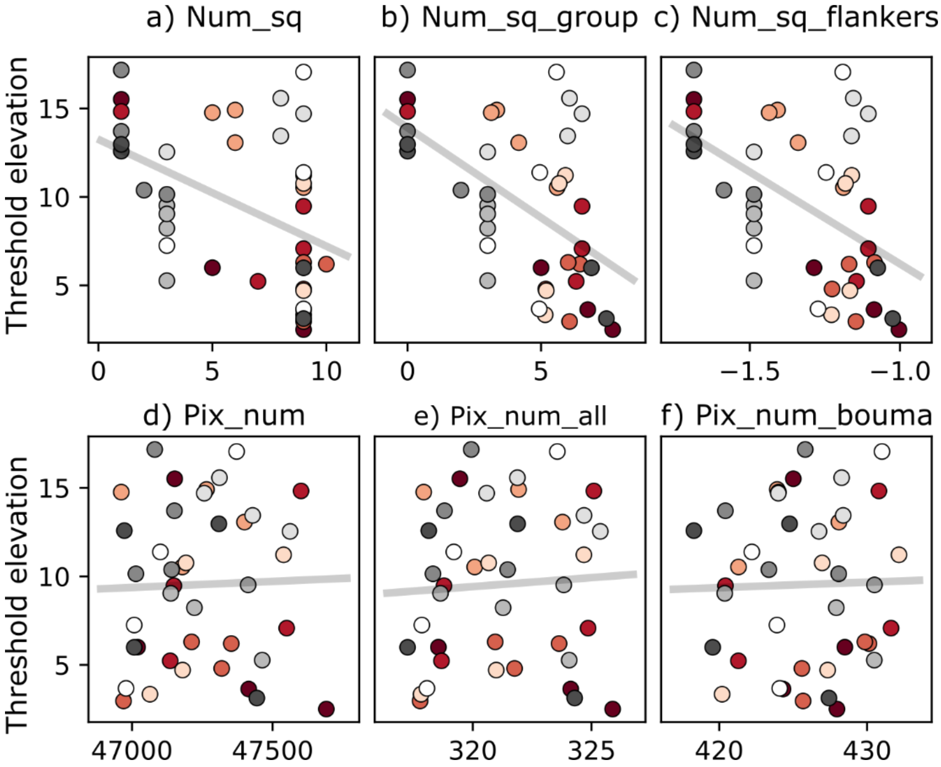


Figure 5. Correlations between model estimates and mean crowding levels of each configuration. The y-axis shows the mean threshold elevation and the x-axis the model estimates for each model (both axes have arbitrary units). Each dot represents each configurations and the color means corresponding Gestalt principles, same as in Figure 3.

Figure 5 shows the correlations between the mean performances across the participants and model predictions. The correlations between crowding level and the number of connected squares and that discounted by the distance showed strong correlation ($r_{\text{num\_sq}}$ (38)= - 0.50, $CI_{95\%}$ = [-0.70, -0.23], $p_{Bonf} < 0.01$; $r_{\text{num\_sq\_group}}$ (38)= - 0.60, $CI_{95\%}$= [-0.75, -0.33], $p_{Bonf} < 0.001$; $r_{\text{num\_sq\_flankers}}$ (38)= - 0.58, $CI_{95\%}$ = [-0.71, -0.23], $p_{Bonf} < 0.001$). However, flanker pixel values, regardless of the local crowding window restriction, showed poor correlation ($r_{\text{pix\_num}}$ (38)= - 0.05, $CI_{95\%}$ = [-0.36, 0.26], p = 0.75; $r_{\text{pix\_num\_all}}$ (38)= -0.02, $CI_{95\%}$ = [-0.33, 0.29], p = 0.90; $r_{\text{pix\_num\_bouma}}$ (38)= - 0.03, $CI_{95\%}$ = [-0.35, 0.29], p = 0.87).

To examine the predictability further, we analyzed the predictability of the models using two methods. First, we used LMMs which had each of the model estimates as the fixed effects. We found that the number of the connected squares and the number of squares with distance discount have a significant effect on the crowding level, but not for the number of pixels. For each LMM, the fixed effect was model estimates for each configuration, and each participant was considered as random intercepts. There were significant fixed effects for the number of directly connected squares based models, but not for the pixel value based models (details in Table 1). Although the effects could only explain 6.0 % of the variances (r_m^2, Num_sq_group; for the other models, see Table 1), it was still better than the pixel estimators (0.0 %, Pix_num_bouma). Note that explained variances including the random intercept across all the models were comparable, 40% - 45% (r_c^2). This result clearly indicates that none of the models can truly explain the crowding and uncrowding, rather there were large performance variences across participants and across configurations. However, the number of directly connected squares and the remaining flankers' distances partly captured the effects.

Table 1. LMM model likelihood test results. Detailed estimates for each model are in Supp. Table xxx.

MODEL	**LIKELIHOOD RATIO TEST**	**SIGNIFICANCE (p)**	**EXPLAINED VARIANCE (r^2)**
NUM_SQ	$\chi^2(1) = 57.077$	$p < 0.001$	$r_m^2 = 0.042,\ \ r_c^2 = 0.433$
NUM_SQ_GROUP	$\chi^2(1) = \ 83.155$	$p < 0.001$	$r_m^2 = 0.060,\ \ r_c^2 = 0.452$
NUM_SQ_FLANKERS	$\chi^2(1) = 76.264$	$p < 0.001$	$r_m^2 = 0.055,\ \ r_c^2 = 0.447$
PIX_NUM	$\chi^2(1) = 0.602$	$p = 0.438$	$r_m^2 = 0.000,\ \ r_c^2 = 0.390$
PIX_N_ALL	$\chi^2(1) = 0.097$	$p = 0.756$	$r_m^2 = 0.000,\ \ r_c^2 = 0.390$
PIX_N_BOUMA	$\chi^2(1) = 0.157$	$p = 0.692$	$r_m^2 = 0.000,\ \ r_c^2 = 0.390$

Next, we tested with leave one out cross validation (LOOCV) method. Hence, here we tested the explained variance of a participants' performance from other remaining participants' performances. We fitted the model estimates to the crowding performance of 19 participants behavioral data. We obtained a r^2 value (explained variance) by using the last participants' data (data points are not included in the linear regression). We repeated the compuation 20 times (for each participant), then averaged the r squared values from 20 iterations to get the final explained variance of each model. As a result, similarly, despite the low correlation, the number of directed squares discounted by their distances could predict the crowding level partially ($r_{\text{LOOCV-num\_sq}}^2$=0.121, $r_{\text{LOOCV-num\_sq\_group}}^2$=0.164, $r_{\text{LOOCV-num\_sq\_flankers}}^2$=0.154), wherewas pixel values could not ($r_{\text{LOOCV-pix\_num}}^2$=0.013, $r_{\text{LOOCV-pix\_n\_all}}^2$=0.013, $r_{\text{LOOCV-pix\_n\_bouma}}^2$=0.015).

Discussion

Visual crowding is a standard vision situation, because visual elements are never preseted alone (reviews: Herzog et al., 2016; Levi, 2008; Pelli & Tillman, 2008; Strasburger, 2020). This has been studied, for a century (e.g., Korte, 1923; Ehlers, 1936; Flom et al., 1963; Bouma, 1970), in an atomic point of view, in which local interactions played major roles. For example, crowding is stronger when features are similar (Kooi et al., 1994; (Bex & Dakin, 2005; Kennedy & Whitaker, 2010; van den Berg et al., 2007). Or flankers only with in Bouma's crowd the target (Bouma, 1970; Bouma, 1973; Levi, 2008; Strasburger et al., 1991; Weymouth, 1958). Therefore, crowding was largely explained by local interaction based models, such as features are pooled (e.g., Dakin et al., 2010; Greenwood et al., 2009, 2017; Parkes et al., 2001; Pelli, 2008; Rosenholtz et al., 2012; Solomon et al., 2004), features may be substituted (Huckauf & Heller, 2002; Strasburger, 2005; Strasburger et al., 1991), or crowding may be mediated by a top-down process (He et al., 1996; Montaser-Kouhsari & Rajimehr, 2005; Tripathy & Cavanagh, 2002; Yeshurun & Rashal, 2010).

However, all these explanations break down when the target is presented with complex, instead of simple flanker configurations (e.g., Livne & Sagi, 2007; Manassi et al., 2016; Põder, 2007; Saarela et al., 2009; Sayim et al., 2010; Yeotikar et al., 2011). Interestingly, flankers faraway from the target can increase (Chanceaux & Grainger, 2013; Manassi et al., 2012; Rosen & Pelli, 2015; Vickery et al., 2009) or decrease (Manassi et al., 2012, 2013, 2015, 2016; Sayim et al., 2010) crowding. Moreover, several models have been proposed for crowing, however, most fail to explain the phenomenon of uncrowding. Doerig and colleagues (2019) compared multiple models of crowding, and showed that only models based on grouping and segmentation processes could explain uncrowding (Choung et al., 2021; Doerig, Bornet, et al., 2020; Doerig, Schmittwilken, et al., 2020).

In this work, we examined three main questions. First of all, we studied whether grouping and segmentation are actual underlying mechanisms that cause uncrowding. Although Manassi et al., (2012) showed that crowding correlates with the participants' subjective ratings about how strongly the vernier target standout from the flanker configuration. However, this was only an indirect measure of grouping and segmentation. We measured five direct and indirect subjective measures to access grouping and segmentation, and found strong correlations. This result provides solid evidence that uncrowding performance is indeed related to grouping and segmentation. Note that there is a dissociation between subjective reports and indirect measures of grouping (e.g., Luna, et al., 2016; Montoro, Villalba-García, Luna & Hinojosa, 2017; Villalba-García, Jimenez, Luna, Hinojosa, & Montoro, 2021; Schmidt & Schmidt, 2013). Moreover, this suggests that the uncrowding paradigm can be used as a quantitative measure for the extent of grouping and segmentation. However, we are aware that correlation does not imply causation (Aldrich, 1995; Tufte, 2006), and the sample size was not large enough to test many hypotheses (I tested 40 conditions with 20 participants, low power and inflated *p-values*). On the other hand, given the high enough inter-rater reliability, these results may be trusted.

Second, we asked whether Gestalt principles can explain (un)crowding by using ten Gestalt principles. We believed that the widely accepted classing Gestalt principles should be able to explain uncrowding (Köhler, 1920; Koffka, 1935; Metzger, 1936, 2006; see reviews by Todorović, 2007; Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al., 2012). However, it was not the case. Instead, (un)crowding depends more on idiosyncratic features of each configuration than a single Gestalt principle. There could be other combinations of Gestalt principles or new principles might explain our dataset better (e.g., common fate: Sekuler & Bennett, 2001; common region: Beck & Palmer, 2002; Palmer, 1992; Palmer & Beck, 2007), further studies are needed.

Lastly, we asked what kinds of computations enable grouping and segmentation. We found mere pixel values (low-level features) cannot explain uncrowding. Rather, the number of directly connected squares was crucial, and it could predict human (un)crowding performance partially. However, we admire that the model could not truly explain the crowding and uncrowding. The results still suggest that perhaps spatial entropy (Altieri et al., 2018) or statistical homogeneity in feature space (Coen-Cagli et al., 2015; Vacher et al., 2019) can result in the grouping of elements.

Crowding is a ubiquitous phenomenon, in which happens overall visual circumstances. It is not restricted to the periphery (also in fovea: Malania, Herzog, & Westheimer, 2007; Sayim, Westheimer, & Herzog, 2010). It occurs in tactile perception (Overvliet & Sayim, 2016) and audition (Oberfeld & Stahn, 2012). It is also found with different stimuli, such as texture (Herrera-Esposito et al., 2020), Gabor (Jastrzębowska et al., 2021; Levi & Carney, 2009; Livne & Sagi, 2007; Põder & Wagemans, 2007; Saarela et al., 2009, 2010), letters (Reuther & Chakravarthi, 2014), shapes (Kimchi & Pirkner, 2015), objects (Wallace & Tjan, 2011), faces (Louie et al., 2007), and tasks (Farzin et al., 2009; Fischer & Whitney, 2011; Yeh et al., 2012). We believe researching perceptual organization with the crowding paradigm provides further insights into the perceptual organization in general, thanks to its ubiquity.

Acknowledgment

We thank GF for the valuable discussions. OHC, MK, MHH were supported by the Swiss National Science Foundation (SNF) 320030_176153 "Basics of visual processing: from elements to figures". ER was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 708007.

References

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 364–376.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated Measures Correlation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00456

Barton, K. (2020). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1--48. https://doi.org/10.18637/jss.v067.i01

Beck, D. M., & Palmer, S. E. (2002). Top-down influences on perceptual grouping. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(5), 1071.

Bex, P. J., & Dakin, S. C. (2005). Spatial interference among moving targets. *Vision Research*, *45*(11), 1385–1398.

Bornet, A., Choung, O.-H., Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (Accepted). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Journal of Vision*.

Bouma, H. (1970). Interaction Effects in Parafoveal Letter Recognition. *Nature*, *226*(5241), 177–178. https://doi.org/10.1038/226177a0

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research*, *13*(4), 767–782.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Casco, C., Campana, G., Han, S., & Guzzon, D. (2009). Psychophysical and electrophysiological evidence of independent facilitation by collinearity and similarity in texture grouping and segmentation. *Vision Research*, *49*(6), 583–593. https://doi.org/10.1016/j.visres.2009.02.004

Chakravarthi, R., & Cavanagh, P. (2009). Recovery of a crowded object by masking the flankers: Determining the locus of feature integration. *Journal of Vision*, *9*(10), 4–4. https://doi.org/10.1167/9.10.4

Chanceaux, M., & Grainger, J. (2013). Constraints on Letter-in-String Identification in Peripheral Vision: Effects of Number of Flankers and Deployment of Attention. *Frontiers in Psychology*, *4*, 119. https://doi.org/10.3389/fpsyg.2013.00119

Choung, O.-H., Bornet, A., Doerig, A., & Herzog, M. H. (2021). Dissecting (un)crowding. *Journal of Vision*.

Claessens, P. M. E., & Wagemans, J. (2005). Perceptual grouping in Gabor lattices: Proximity and alignment. *Perception & Psychophysics*, *67*(8), 1446–1459. https://doi.org/10.3758/BF03193649

Claessens, P. M. E., & Wagemans, J. (2008). A Bayesian framework for cue integration in multistable grouping: Proximity, collinearity, and orientation priors in zigzag lattices. *Journal of Vision*, *8*(7), 33–33. https://doi.org/10.1167/8.7.33

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, *18*, 1648.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Dakin, S. C., Cass, J., Greenwood, J. A., & Bex, P. J. (2010). Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli. *Journal of Vision*, *10*(10), 14–14.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020a). Crowding reveals fundamental differences in local vs. Global processing in humans and machines. *Vision Research*, *167*, 39–45.

Doerig, A., Bornet, A., Choung, O.-H., & Herzog, M. H. (2020b). Crowding reveals fundamental differences in local vs. Global processing in humans and machines. *Vision Research*, *167*, 39–45. https://doi.org/10.1016/j.visres.2019.12.006

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. https://doi.org/10.1371/journal.pcbi.1006580

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.

Ehlers, H. (1936). V: The movements of the eyes during reading. *Acta Ophthalmologica*, *14*(1-2), 56–63.

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), 483–504. https://doi.org/10.1037/rev0000070

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6), 16–16. https://doi.org/10.1167/11.6.16

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, *17*(1), 3.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Preprint ArXiv:1811.12231*.

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.

Greenwood, J. A., Szinte, M., Sayim, B., & Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceedings of the National Academy of Sciences*, *114*(17), E3573–E3582. https://doi.org/10.1073/pnas.1615504114

Hadad, B., Maurer, D., & Lewis, T. L. (2010). The effects of spatial proximity and collinearity on contour integration in adults and children. *Vision Research*, *50*(8), 772–778. https://doi.org/10.1016/j.visres.2010.01.021

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, *383*(6598), 334–337. https://doi.org/10.1038/383334a0

Herzog, M. H., Thunell, E., & Ögmen, H. (2016). Putting low-level vision into global context: Why vision cannot be reduced to basic circuits. *Vision Research*, *126*, 9–18. https://doi.org/10.1016/j.visres.2015.09.009

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Huckauf, A., & Heller, D. (2002). What various kinds of errors tell us about lateral masking effects. *Visual Cognition*, *9*(7), 889–910.

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Kennedy, G. J., & Whitaker, D. (2010). The chromatic selectivity of visual crowding. *Journal of Vision*, *10*(6), 15–15.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

Koffka, K. (1935). Principles of Gestalt Psychology, International Library of Psychology. *Philosophy and Scientific Method*, *32*(8).

Köhler, W. (1920). *Die physischen Gestalten in Ruhe und im stationaren Eine natur-philosophische Untersuchung [The physical Gestalten at rest and in steady state].* Springer.

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

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

Kovács, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(16), 7495–7497.

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. https://doi.org/10.1016/j.visres.2007.12.009

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision*, *7*(2), 4. https://doi.org/10.1167/7.2.4

Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision*, *15*(8), 16–16. https://doi.org/10.1167/15.8.16

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

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. https://doi.org/10.1167/12.10.13

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision*, *13*(13), 10–10. https://doi.org/10.1167/13.13.10

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Metzger, W. (1936). *Gesetze des Sehens [Laws of seeing].* (Frankfurt am Main, Germany).

Metzger, W., Spillmann, L. T., Lehar, S. T., Stromeyer, M. T., & Wertheimer, M. T. (2006). *Laws of seeing.* Mit Press.

Mohr, D. L., & Marcon, R. A. (2005). Testing for a ‘within-subjects’ association in repeated measures data. *Journal of Nonparametric Statistics*, *17*(3), 347–363. https://doi.org/10.1080/10485250500038694

Montaser-Kouhsari, L., & Rajimehr, R. (2005). Subliminal attentional modulation in crowding condition. *Vision Research*, *45*(7), 839–844. https://doi.org/10.1016/j.visres.2004.10.020

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Oberfeld, D., & Stahn, P. (2012). Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PloS One*, *7*(10), e48054.

Overvliet, K. E., & Sayim, B. (2016). Perceptual grouping determines haptic contextual modulation. *Vision Research*, *126*, 52–58.

Palmer, S. E. (1992). Common region: A new principle of perceptual grouping. *Cognitive Psychology*, *24*(3), 436–447.

Palmer, S. E., & Beck, D. M. (2007). The repetition discrimination task: An objective method for studying perceptual grouping. *Perception & Psychophysics*, *69*(1), 68–78.

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. https://doi.org/10.1038/89532

Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology*, *18*(4), 445–451.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135. https://doi.org/10.1038/nn.2187

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

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

R Core Team. (2019). *R: A Language and Environment for Statistical Computing* (Vienna, Austria). R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Evanston, Illinois). Northwestern University. https://CRAN.R-project.org/package=psych

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

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.

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

Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, *9*(2), 5–5. https://doi.org/10.1167/9.2.5

Saarela, T. P., Westheimer, G., & Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *Journal of Vision*, *10*(10), 17–17. https://doi.org/10.1167/10.10.17

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 3856–3866.

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in vernier acuity. *Journal of Vision*, *8*(8), 12–12. https://doi.org/10.1167/8.8.12

Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt Factors Modulate Basic Spatial Vision. *Psychological Science*, *21*(5), 641–644. https://doi.org/10.1177/0956797610368811

Sekuler, A. B., & Bennett, P. J. (2001). Generalized common fate: Grouping by common luminance changes. *Psychological Science*, *12*(6), 437–444.

Solomon, J. A., Felisberti, F. M., & Morgan, M. J. (2004). Crowding and the tilt illusion: Toward a unified account. *Journal of Vision*, *4*(6), 9–9. https://doi.org/10.1167/4.6.9

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Strasburger, H. (2005). Unfocussed spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision*, *5*(11), 8–8. https://doi.org/10.1167/5.11.8

Strasburger, H. (2020). Seven Myths on Crowding and Peripheral Vision. *I-Perception*, *11*(3), 2041669520913052. https://doi.org/10.1177/2041669520913052

Strasburger, H., Harvey, L. O., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Perception & Psychophysics*, *49*(6), 495–508.

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient Estimates on Probability Functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787. https://doi.org/10.1121/1.1910407

Todorović, D. (2007). W. Metzger, Laws of Seeing. *Gestalt Theory*, *29*(2), 176.

Tripathy, S. P., & Cavanagh, P. (2002). The extent of crowding in peripheral vision does not scale with target size. *Vision Research*, *42*(20), 2357–2369. https://doi.org/10.1016/S0042-6989(02)00197-9

Tufte, E. R. (2006). *Beautiful evidence* (Vol. 1). Graphics Press Cheshire, CT.

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, *7*(2), 14–14.

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. https://doi.org/10.1167/9.2.12

von Ehrenfels, C. (1890). Über Gestaltqualitäten_About gestalt qualities. *Vierteljahrsschrift Für Wissenschaftliche Philosophie*, *14*, 249–292.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218.

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

Wertheimer, M. (1912). Experimentelle studien uber das sehen von bewegung. *Zeitschrift Fur Psychologie*, *61*.

Wertheimer, M. (1922). Untersuchungen zur Lehre von der Gestalt I: Prinzipielle Bemerkungen [Investigations in Gestalt theory: I. The general theoretical situation]. *Psychologische Forschung*, *1*(1), 47–58.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. [Investigations in Gestalt Theory: II. Laws of organization in perceptual forms]. *Psychologische Forschung*, *4*, 301–350.

Weymouth, F. W. (1958). Visual sensory units and the minimal angle of resolution. *American Journal of Ophthalmology*, *46*(1), 102–113.

Yeotikar, N. S., Khuu, S. K., Asper, L. J., & Suttle, C. M. (2011). Configuration specificity of crowding in peripheral vision. *Vision Research*, *51*(11), 1239–1248. https://doi.org/10.1016/j.visres.2011.03.016

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. https://doi.org/10.1167/10.10.16

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21–29. https://doi.org/10.22237/jmasm/1177992180

Supplementary Materials

Methods

Statistics of Repeated measures Spearman's Rank correlation

We have twenty participants with forty different conditions. Therefore, the usual correlation analysis is not applicable as the data violate the assumption of independence. We largely referred Mohr and Marcon (2005) and Bakdash and Marusich (2017) works and used the randomization test (Edgington, & Onghena, 2007) to acquire the significance of the correlation. Spearman's rank correlation was utilized to account for the ordinal data. All the analysis codes are available online github.com/Ohyeon5/#####.

Significance test (p-values)

The significance of the correlation was tested with the randomization test. The procedure was as follows.

1. Compute the Spearman correlation per participant (with 40 conditions)
2. Average the Spearman correlation coefficient across participants
3. Randomly permute the rank within each participant (1,000,000 times)
4. Compute the mean Spearman correlation coefficient of each randomized set
5. Count the number of coefficients larger than the originally acquired coefficient
6. Draw a histogram and fit a gaussian function to get a probability distribution function (p-values)

Confidence interval (CI)

We followed the bootstraping method utilized in Bakdash and Marusich, 2017 (rmcorr) to obtain 95% CIs for mean Spearman's rank correlation coefficients. There are numerous analytic methods to estimate CIs (e.g., Fisher, 1921; Woods, 2007), but the methods are not appropriate for our dataset as parametric asumptions are needed. Instead, we used the bootstraping which does not require distributional assumptions and uses resampling to estimate parameter accuracy (Efron & Tibshirani, 1994).

Here, we resampled the data within each participant for 10,000 times (bootstrap samples), and computed the mean Spearman correlation coefficient of each resampled data. Then, we constructed an empirical sampling distribution of the mean Spearman's Rank correlation coefficients. We obtained the CIs from the empirical sampling distribution (CI_{emp}) and with the percentile bootstrap method ($CI_{percent}$; Efron & Tibshirani, 1994). As CI_{emp} and $CI_{percent}$ were comparable, we only presented $CI_{percent}$.

References

Edgington, E. & Onghena, P. (2007) Randomization tests, New York, Chapman & Hall.

Ruscio, J. (2008). Constructing confidence intervals for Spearman's rank correlation with ordinal data: a simulation study comparing analytic and bootstrap methods. *Journal of Modern Applied Statistical Methods*, *7*(2), 7.

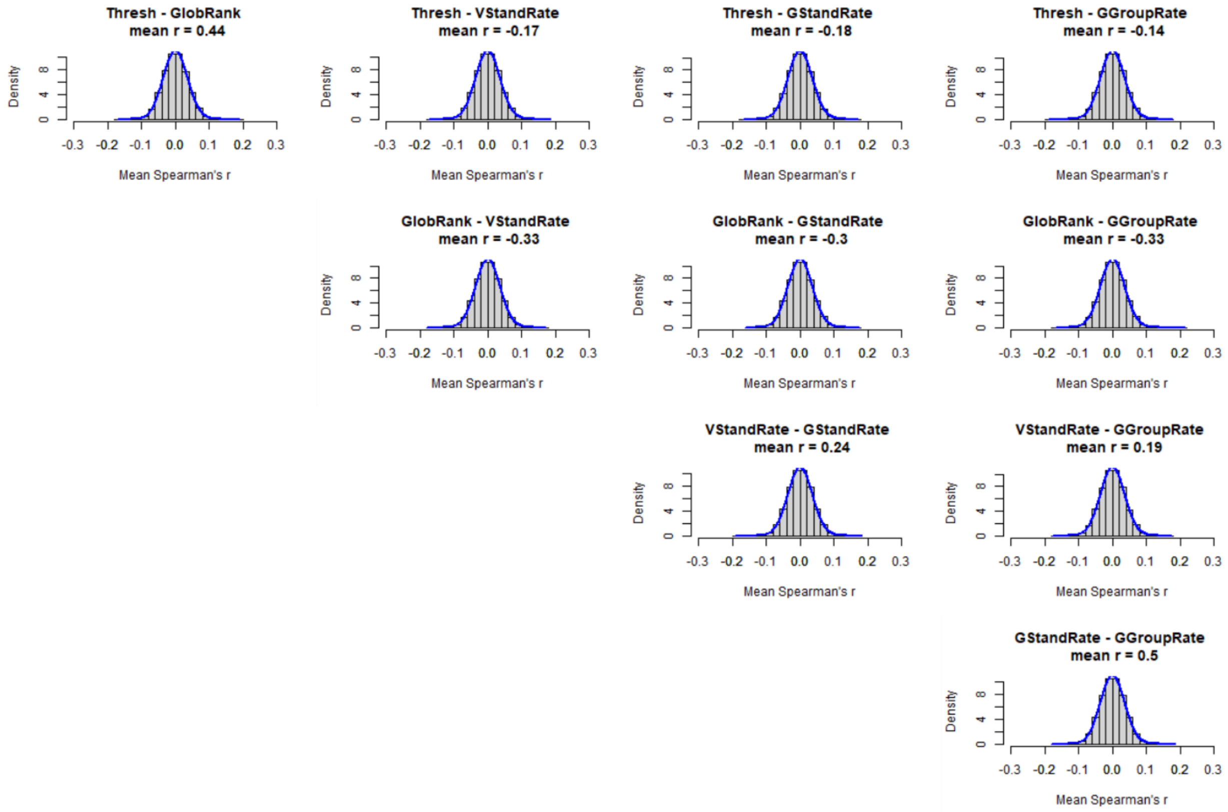
Supplementary Tables

Parameter estimates of Linear Mixed Effects Models (LMMs)

Table 1. Estimates from the linear mixed-effects model of the VCrowd task, with the Gestalt principles as predictors and individual participants and flanker configurations as random intercepts.

Fixed Effects	β estimate	β standard error	t-value
(Intercept)	2.043	0.256	7.985
Close-SymmClose	-0.124	0.256	-0.483
Close-IrreCont	-0.413	0.256	-1.612
Close-CollCont	-0.754	0.256	-2.941
Close-Randw	-1.612	0.314	-3.701
Close-Randwo	0.056	0.314	0.178
Close-Repeat	-0.128	0.256	-0.501
Close-RepeatDia	-0.660	0.256	-2.573
Close-Symm1	-0.803	0.222	-3.617
Close-Symm2	-1.019	0.256	-3.973

Supplementary Figures



Supp. Figure 4. Spearman's Rank coefficient randomization test histogram.

Chapter 4 Conclusions

4.1 Summary

Visual processing is thought to be feedforward, hierarchical, localized, and based on local interactions (Hubel et al., 1978; Hubel & Wiesel, 1959, 1962; Riesenhuber & Poggio, 1999). However, perception is global and recursive processing. Elements are integrated across space and time, forming the fundamental units of visual processing (reviews: Todorović, 2007; Wagemans, 2015). In Fig. 1.1.1a, we perceive a fence and a house rather than a mesh pattern drawn on the house. This occurs despite parts of the mesh fence being positioned closer to the house than the other parts. Note that it is only valid for the 2D projected cases (e.g., in a photo), in the real-world situation, the depth makes parts of the mesh fence to be positioned closer to each other. Here, the perception occurred in two steps. First, elements (e.g., part of mesh fence) are integrated into an object (e.g., fence). Second, parts are perceived based on their membership. Vision is a dynamic process requiring both feedforward and feedback processing to integrate elements. The traditional model of vision cannot explain this. In this thesis, I used non-retinotopic processing for understanding spatiotemporal integration and visual (un)crowding for spatial integration.

Motion perception of parts of the object occurs relative to the objects that parts belong to. Thus motion perception is non-retinotopic. For example, as in Fig. 1.2.2, a reflector on a bicycle wheel is perceived to rotate, despite its cycloidal motion in the retinotopic coordinate. This is because the reflector motion is perceived relative to the bike. The horizontal motion of the bike is subtracted from the reflector's retinotopic cycloidal motion. In the Ternus-Pikler display (TPD), dot motion is perceived relative to the disk motion. The TPD is the perfect probe to capture the non-retinotopic process, because it can pit retinotopic versus non-retinotopic information. The TPD can also be generalized to other psychophysical experiments, such as visual search and motion adaptation (Boi et al., 2009).

In chapter 2, I used the TPD to study the properties of spatiotemporal integration. First of all, I showed that non-retinotopic motion perception in the TPD is not merely attentional tracking. I enhanced the saliency of the target in retinotopic motion conditions. Even though the target was clearly different from the other disks and hence well traceable, the performance improvement was not to the extent expected from a tracking mechanism. The results provide further evidence that non-retinotopic perception originates from largely inbuilt spatiotemporal integration. Next, I showed that invisible retinotopic interpretation can affect visible non-retinotopic perception. When pitting dot rotations of retinotopic interpretation against non-retinotopic interpretation, performance was affected by the invisible retinotopic process. The interference of invisible stimuli is consistent with previous research in masked priming, where an invisible prime can speed up or delay responses to subsequently presented targets. This effect is typically explained by a short-lived pre-activation of the motor system (e.g., Klotz & Neumann, 1999; Klotz & Wolff, 1995). However, unexpectedly, in TPD, only incongruent retinotopic rotation affected the conscious perception, but not by congruent information. This cannot be explained by other mechanisms, for example, alternating boundary ownership in ambiguous figures (Layton et al., 2012; Zhaoping, 2005; Zhou et al., 2000). I suggest that there are multiple channels to process conscious motion (details see chapter 2.2). However, very little is known about the neurophysiology of non-retinotopic processing. Thunell and colleagues (2016) found that the average blood-oxygen-level (BOLD) activation in early visual areas such as V1, V2, and V3 reflects the retinotopic properties but not non-retinotopic properties in TPD. In the human motion processing complex (hMT+), BOLD signal encoded both retinotopic and non-retinotopic properties. This fMRI result suggests that hMT+ may be the locus of unconscious retinotopic influences on conscious non-retinotopic percept. That is, hMT+ as the first locus of encoding non-retinotopic percept preserves both retinotopic and non-retinotopic interpretations. Thus, dynamic interactions between retinotopic and non-retinotopic interpretations are processed in hMT+ to have conscious motion perception. Further neurophysiological studies are needed.

Visual elements are never presented alone in everyday vision. Visual crowding occurs when a target is presented in clutter, a situation in which target perception deteriorates. Interestingly, the target can be released from crowding when more flankers are presented, depending on flanker configurations. For example, in Fig. 1.3.2, performance deteriorates when the Vernier is presented with a square flanker (classic crowding). When more squares are presented, performance improves (a release of crowding; uncrowding). Uncrowding highly depends on the global configuration of flankers (Fig. 1.3.2 right panel). In uncrowding, the target and flankers are integrated into different groups, which serve as reference frames for perception. Thus, the target is perceived relative to its own group, and the square is perceived relative to the squares group. Uncrowding is not restricted to the periphery (also in fovea: Malania, Herzog, & Westheimer, 2007; Sayim, Westheimer, & Herzog, 2010). It occurs in tactile perception (Overvliet & Sayim, 2016) and audition (Oberfeld & Stahn, 2012). It is also found with different stimuli, such as texture (Herrera-Esposito et al., 2020), Gabor (Jastrzębowska et al., 2021; Levi & Carney, 2009; Livne & Sagi, 2007; Põder & Wagemans, 2007; Saarela et al., 2009, 2010), letters (Reuther & Chakravarthi, 2014), shapes (Kimchi & Pirkner, 2015), objects (Wallace & Tjan, 2011), faces (Louie et al., 2007), and different tasks (Farzin et al., 2009; Fischer & Whitney, 2011; Yeh et al., 2012). Questions remain about whether more complex feedforward processes can explain uncrowding and what matters in uncrowding.

In chapter 3, I used visual crowding to answer the above questions. First, in chapter 3.1, using state-of-the-art ffCNNs pre-trained with the large-scale dataset (ImageNet: Deng et al., 2009; or shape-biased ImageNet: Geirhos et al., 2018), I showed that feedforward processing (as in the classic model of vision) cannot reproduce uncrowding. Although pre-trained ffCNNs may account for global shape information (Geirhos et al., 2018), they could not automatically produce grouping and segmentation processes. The result indicates that grouping and segmentation cannot occur by feedforward processes. In my other contribution (Bornet et al., 2021), I also showed that a more complex pooling process, texture tiling model (TTM; Rosenholtz et al., 2019), cannot explain uncrowding either. I propose that recurrent, global grouping and segmentation are crucial to explaining how the brain deals with global configurations. In a follow-up study, Doerig and colleagues (Doerig, Schmittwilken, et al., 2020) showed recurrent models with grouping and segmentation such as Capsule networks (Sabour et al., 2017) can explain uncrowding. Moreover, in chapter 3.3, I compared three models (Capsule networks, the Laminart model, and TTM) on different computation bases, and again confirmed that two-stage models, which group and segment based on objects, are needed. One may suggest that feedforward CNNs which do per object segmentations, such as instance segmentation (K. He et al., 2017), can explain (un)crowding, thanks to their explicit segmentation process. However, it is important to note that instance segmentation models are, in principle, two-stage models. The models first detect the object using a bounding box, and then classify each pixel within the bounding box for segmentation. Therefore, again this emphasizes the two-stage models, which group and segment based on objects, are needed.

Then, I analyzed the configurations of crowding and uncrowding to understand what matters in (un)crowding. In chapter 3.2, I used the classic crowding stimuli to examine how target-related information is integrated within a crowding window. In the classic crowding stimulus, target and flankers are in the same perceptual group, so all the elements interfere (no grouping and segmentation). I tested the Vernier offset discrimination task with various combinations of lines and Verniers as flankers. The results are mixed, but one clear observation is that crowding cannot be explained by a single process. Importantly, crowding is not simply averaging elements within a receptive field (pooling; e.g., Dakin et al., 2010; Freeman et al., 2012; Greenwood et al., 2009; Parkes et al., 2001). Crowding is also not simply substitution (e.g., Chung et al., 2001; Ester et al., 2015; Huckauf & Heller, 2002; Strasburger, 2005; Strasburger et al., 1991).

Next, in chapter 3.3, I dissected the uncrowding stimuli and examined to what extent low-level features matter. I systematically dissected a stimulus configuration to test this. When only subparts of the flankers were presented, there was always strong crowding, indicating that configural aspects, such as the Gestalt principle of Prägnanz, seem to matter. I showed that uncrowding cannot be explained by low-level interactions, such as line-line detector inhibitions in the surrounding suppression or divisive normalization model (Carandini & Heeger, 2012; Coen-Cagli et al., 2015) or contour-contour integration (illusory contour: Francis et al., 2017; Clarke, Herzog, et al., 2014; Doerig et al., 2019). Then, I showed that low-level features, such as orientations or radial-tangential anisotropy, have minor impact on uncrowding.

Next, in chapter 3.4, I tested whether basic Gestalt principles can explain uncrowding (Köhler, 1920; Koffka, 1935; Metzger, 1936, 2006; see reviews by Todorović, 2007; Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al., 2012). Gestalt principles could not fully explain (un)crowding. For the same Gestalt principle, some flanker configurations showed good performance, whereas the others did not. This result suggests that uncrowding and grouping depend more on idiosyncratic features of each configuration than a single Gestalt principle. However, shape similarity is not the strongest grouping principle, and there could be other combinations of Gestalt principles or new principles might explain our dataset better (e.g., common fate: Sekuler & Bennett, 2001; common region: Beck & Palmer, 2002; Palmer, 1992; Palmer & Beck, 2007).

Finally, I studied whether grouping and segmentation are actual underlying mechanisms that cause uncrowding. To make a long story short, I found strong correlations between uncrowding and grouping and segmentation. Manassi and colleagues (2013) showed that the central square is harder to discriminate (highly crowded; aspect ratio discrimination task in chapter 3.3) when increasing the number of squares, and the crowding of the squares was strongly correlated with the uncrowding of the Vernier target. The authors suggested that the strong square crowding is because elements are first grouped together, and only elements within a group interfere. I replicated the results in chapter 3.3. In addition, Manassi et al., (2012) showed that crowding correlates with the participants' subjective ratings about how strongly the vernier target stand-out from the flanker configuration. However, this was only an indirect measure of grouping and segmentation. In chapter 3.4, I measured five direct and indirect subjective measures to access grouping and segmentation. I found a strong correlation between the Vernier offset threshold and the subjective grouping measures. This result provides solid evidence that uncrowding performance is indeed related to grouping and segmentation. Moreover, this suggests that the uncrowding paradigm can be used as a quantitative measure for the extent of grouping and segmentation. However, I am aware that correlation does not imply causation (Aldrich, 1995; Tufte, 2006), and the sample size was not large enough to test many hypotheses (I tested 40 conditions with 20 participants, low power and inflated *p-values*).

However, the neural substrates of uncrowding remain unclear. A series of literature has shown that BOLD signals are attenuated in the crowding conditions in multiple stages of visual processing, such as in V1, V2, V3, V4, and lateral occipital complex (LOC) (Anderson et al., 2012; Bi et al., 2009; Fang & He, 2008; Millin et al., 2014). Very little is known about uncrowding. Jastrzębowska and colleagues (2021) observed interesting BOLD signal changes across the crowding and uncrowding conditions. As expected, the BOLD signal was attenuated in crowding conditions compared to the only target conditions in V1, V2, V3, V4, and LOC. However, unexpectedly, the BOLD signal was further suppressed in uncrowding conditions. Such further suppressed BOLD responses held in all regions of interest (ROIs), V1, V2, V3, V4, and LOC. These results indicate complex and recursive interactions among different areas. The authors further utilized dynamic causal modeling (DCM) to determine how five ROIs interact with each other, and showed that the recursive model, which includes both feedforward and feedback interactions among V2 to V4 and LOC, explains the uncrowding condition the best. Therefore, Jastrzębowska and colleagues showed complex recursive interactions among different areas are needed to explain uncrowding, but further studies are needed to understand which exact interactions are required.

Taken together, spatial and temporal integration is rather a complex inbuilt automated mechanism, and integration occurs across the whole visual field. Moreover, perception is best explained by two-stage processes, which include grouping and segmentation. Thereby, I suggest, to better understand the integration across space and time, we need to consider a model that groups elements by multiple processes and recursively segments other groups.

4.2 Future development

My work opens directions for future studies, both in experimental and modeling approaches. First, in the experimental approach, more complex flanker configurations need to be tested. For example, the flanker configurations using other Gestalt principles, including old and basic principles, new principles, and systematic combinations of the principles, could be tested. Otherwise, object-like configurations (e.g., animals, faces, plants) could be used (Neri, 2014, 2017) to test the top-down impacts on grouping and segmentation. Another possibility is to use temporal crowding paradigms, as suggested by Yeshurun and colleagues (2015). The authors showed that the target perception is impaired when flankers are presented in a different time frame. We may observe uncrowding performance with such a temporal crowding paradigm. Moreover, temporal crowding may even be tested with the TPD display (visual research task could be done in TPD; Boi et al., 2009). The observations from Vernier stimuli should be generalized to crowding paradigms or with naturalistic stimuli (e.g., Livne & Sagi, 2007; Manassi et al., 2016; Põder, 2007; Saarela et al., 2009; Sayim et al., 2010; Yeotikar et al., 2011). Second, modeling insights should be acquired from these rich experimental data. For example, in chapter 3.4, I observed that the number of directly connected squares predicts crowding performance well, suggesting that grouping may occur when a feature is closely and repeatedly presented. This observation gives good insights that perhaps spatial entropy (Altieri et al., 2018) or statistical homogeneity in feature space (Coen-Cagli et al., 2015; Vacher et al., 2019) can result in the grouping of elements. These measures may serve as a constraint for the model to group and segment.

To this end, I believe further experimental observations and modeling works on grouping and segmentation will help us understand human vision better. I suggest that this will further facilitate the path to a more robust and reliable model of vision.

References

Adler, F. H., & Moses, R. A. (1965). *Physiology of the eye: Clinical application*. Mosby.

Agaoglu, M. N., Clarke, A. M., Herzog, M. H., & Öğmen, H. (2016). Motion-based nearest vector metric for reference frame selection in the perception of motion. *Journal of Vision*, *16*(7), 14. https://doi.org/10.1167/16.7.14

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 364–376.

Altieri, L., Cocchi, D., & Roli, G. (2018). A new approach to spatial entropy measures. *Environmental and Ecological Statistics*, *25*(1), 95–110.

Anderson, E. J., Dakin, S. C., Schwarzkopf, D. S., Rees, G., & Greenwood, J. A. (2012). The neural correlates of crowding-induced changes in appearance. *Current Biology*, *22*(13), 1199–1206.

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

Bach, M. (1996). The Freiburg Visual Acuity Test—Automatic Measurement of Visual Acuity. *Optometry and Vision Science*, *73*(1), 49–53.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13–13. https://doi.org/10.1167/9.12.13

Barton, K. (2020). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1--48. https://doi.org/10.18637/jss.v067.i01

Beck, D. M., & Palmer, S. E. (2002). Top-down influences on perceptual grouping. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(5), 1071.

Bergen, J. R., & Landy, M. S. (1991). Computational modeling of visual texture segregation. *Computational Models of Visual Processing*, *17*, 253–271.

Bernard, J.-B., & Chung, S. T. L. (2011). The dependence of crowding on flanker complexity and target–flanker similarity. *Journal of Vision*, *11*(8), 1–1. https://doi.org/10.1167/11.8.1

Bex, P. J., & Dakin, S. C. (2005). Spatial interference among moving targets. *Vision Research*, *45*(11), 1385–1398.

Bi, T., Cai, P., Zhou, T., & Fang, F. (2009). The effect of crowding on orientation-selective adaptation in human early visual cortex. *Journal of Vision*, *9*(11), 13–13.

Boi, M., Öğmen, H., Krummenacher, J., Otto, T. U., & Herzog, M. H. (2009). A (fascinating) litmus test for human retino- vs. Non-retinotopic processing. *Journal of Vision*, *9*(13), 5–5. https://doi.org/10.1167/9.13.5

Bornet, A., Choung, O.-H., Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (2021). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Journal of Vision*, *21*(12), 10. https://doi.org/10.1167/jov.21.12.10

Bouma, H. (1970). Interaction Effects in Parafoveal Letter Recognition. *Nature*, *226*(5241), 177–178. https://doi.org/10.1038/226177a0

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research*, *13*(4), 767–782.

Bouma, H., & Andriessen, J. J. (1968). Perceived orientation of isolated line segments. *Vision Research*, *8*(5), 493–507.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brendel, W., & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. *ArXiv Preprint ArXiv:1904.00760*.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Choung, O.-H., Bornet, A., Doerig, A., & Herzog, M. H. (2021). Dissecting (un)crowding. *Journal of Vision*, *21*(10), 1–20. https://doi.org/10.1167/jov.21.10.10

Choung, O.-H., Rashal, E., & Herzog, M. H. (2019). Basic gestalt laws cannot explain uncrowding. *Perception*, *48*(CONF), 28–28.

Chung, S. T. L. (2013). Cortical Reorganization after Long-Term Adaptation to Retinal Lesions in Humans. *Journal of Neuroscience*, *33*(46), 18080–18086. https://doi.org/10.1523/JNEUROSCI.2764-13.2013

Chung, S. T. L., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research*, *41*(14), 1833–1850. https://doi.org/10.1016/S0042-6989(01)00071-2

Clarke, A. M., Öğmen, H., & Herzog, M. H. (2016). A computational model for reference-frame synthesis with applications to motion perception. *Vision Research*, *126*, 242–253. https://doi.org/10.1016/j.visres.2015.08.018

Coates, D. R., Bernard, J.-B., & Chung, S. T. L. (2019). Feature contingencies when reading letter strings. *Vision Research*, *156*, 84–95. https://doi.org/10.1016/j.visres.2019.01.005

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, *18*, 1648.

Dakin, S. C., Cass, J., Greenwood, J. A., & Bex, P. J. (2010). Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli. *Journal of Vision*, *10*(10), 14–14.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

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. https://doi.org/10.1016/j.visres.2019.12.006

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. https://doi.org/10.1371/journal.pcbi.1006580

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.

Duncker, K. (1929). Über induzierte Bewegung. *Psychologische Forschung*, *12*(1), 180–259. https://doi.org/10.1007/BF02409210

Ehlers, H. (1936). V: The movements of the eyes during reading. *Acta Ophthalmologica*, *14*(1-2), 56–63.

Ehlers, H. (1953). CLINICAL TESTING OF VISUAL ACUITY. *A.M.A. Archives of Ophthalmology*, *49*(4), 431–434. https://doi.org/10.1001/archopht.1953.00920020441007

Ester, E. F., Klee, D., & Awh, E. (2014). Visual crowding cannot be wholly explained by feature pooling. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1022.

Ester, E. F., Zilber, E., & Serences, J. T. (2015). Substitution and pooling in visual crowding induced by similar and dissimilar distractors. *Journal of Vision*, *15*(1), 4–4. https://doi.org/10.1167/15.1.4

Fabre-Thorpe, M. (2011). The Characteristics and Limits of Rapid Visual Categorization. *Frontiers in Psychology*, *2*, 243. https://doi.org/10.3389/fpsyg.2011.00243

Fang, F., & He, S. (2008). Crowding alters the spatial distribution of attention modulation in human primary visual cortex. *Journal of Vision*, *8*(9), 6–6.

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision*, *9*(6), 18.1-15. https://doi.org/10.1167/9.6.18

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field." *Vision Research*, *33*(2), 173–193. https://doi.org/10.1016/0042-6989(93)90156-Q

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology*, *106*(3), 1389–1398.

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review*, *124*(4), 483–504. https://doi.org/10.1037/rev0000070

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Attention, Perception, & Psychophysics*, *74*(2), 379–396. https://doi.org/10.3758/s13414-011-0229-0

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201. https://doi.org/10.1038/nn.2889

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Preprint ArXiv:1811.12231*.

Gheri, C., & Baldassi, S. (2008). Non-linear integration of crowded orientation signals. *Vision Research*, *48*(22), 2352–2358. https://doi.org/10.1016/j.visres.2008.07.022

Gheri, C., Morgan, M. J., & Solomon, J. A. (2007). The relationship between search efficiency and crowding. *Perception*, *36*(12), 1779–1787.

Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., & He, K. (2018). *Detectron*.

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.

Greenwood, J. A., Szinte, M., Sayim, B., & Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceedings of the National Academy of Sciences*, *114*(17), E3573–E3582. https://doi.org/10.1073/pnas.1615504114

Hafiz, A. M., & Bhat, G. M. (2020). A survey on instance segmentation: State of the art. *International Journal of Multimedia Information Retrieval*, 1–19.

Harrison, W. J., & Bex, P. J. (2015). *Current Biology*, *25*(24), 3213–3219. https://doi.org/10.1016/j.cub.2015.10.052

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, S., Cavanagh, P., & Intriligator, J. (1996). Attentional resolution and the locus of visual awareness. *Nature*, *383*(6598), 334–337. https://doi.org/10.1038/383334a0

Herrera-Esposito, D., Coen-Cagli, R., & Gomez-Sena, L. (2020). Flexible contextual modulation of naturalistic texture perception in peripheral vision. *BioRxiv*, 2020.01.24.918813. https://doi.org/10.1101/2020.01.24.918813

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.

Herzog, M. H., Thunell, E., & Ögmen, H. (2016). Putting low-level vision into global context: Why vision cannot be reduced to basic circuits. *Vision Research*, *126*, 9–18. https://doi.org/10.1016/j.visres.2015.09.009

Hinton, G. E. (1981, January 1). *A Parallel Computation that Assigns Canonical Object-Based Frames of Reference*. IJCAI. https://openreview.net/forum?id=SkWXH7fO-r

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*, 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Hubel, D. H., Wiesel, T. N., & Stryker, M. P. (1978). Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, *177*(3), 361–379.

Huckauf, A., & Heller, D. (2002). What various kinds of errors tell us about lateral masking effects. *Visual Cognition*, *9*(7), 889–910.

Jastrzębowska, M. A., Chicherov, V., Draganski, B., & Herzog, M. H. (2021). Unraveling brain interactions in vision: The example of crowding. *NeuroImage*, *240*, 118390. https://doi.org/10.1016/j.neuroimage.2021.118390

Johansson, G. (1950). Configurations in the perception of velocity. *Acta Psychologica*, *7*, 25–79. https://doi.org/10.1016/0001-6918(50)90003-5

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. https://doi.org/10.3758/BF03212378

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Kennedy, G. J., & Whitaker, D. (2010). The chromatic selectivity of visual crowding. *Journal of Vision*, *10*(6), 15–15.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*, 133504.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. *ArXiv Preprint ArXiv:1906.01558*.

Kim, T., Bair, W., & Pasupathy, A. (2019). Neural coding for shape and texture in macaque area V4. *Journal of Neuroscience*, *39*(24), 4760–4774.

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.

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?*

Klotz, W., & Neumann, O. (1999). Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 976–992. https://doi.org/10.1037/0096-1523.25.4.976

Klotz, W., & Wolff, P. (1995). The effect of a masked stimulus on the response to the masking stimulus. *Psychological Research*, *58*(2), 92–101. https://doi.org/10.1007/BF00571098

Koffka, K. (1935). Principles of Gestalt Psychology, International Library of Psychology. *Philosophy and Scientific Method*, *32*(8).

Köhler, W. (1920). *Die physischen Gestalten in Ruhe und im stationaren Eine natur-philosophische Untersuchung [The physical Gestalten at rest and in steady state].* Springer.

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

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

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*(1), 26–49. https://doi.org/10.1016/j.tics.2012.10.011

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

Kwon, M., Bao, P., Millin, R., & Tjan, B. S. (2014). Radial-tangential anisotropy of crowding in the early visual areas. *Journal of Neurophysiology*, *112*(10), 2413–2422. https://doi.org/10.1152/jn.00476.2014

Kwon, M., & Liu, R. (2019). Linkage between retinal ganglion cell density and the nonuniform spatial integration across the visual field. *Proceedings of the National Academy of Sciences*, *116*(9), 3827–3836. https://doi.org/10.1073/pnas.1817076116

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lauffs, M. M. (2017). *From retinotopic processing to nonretinotopic representation* [EPFL]. https://doi.org/10.5075/epfl-thesis-8126

Lauffs, M. M., Choung, O.-H., Öğmen, H., & Herzog, M. H. (2018). Unconscious retinotopic motion processing affects non-retinotopic motion perception. *Consciousness and Cognition*. https://doi.org/10.1016/j.concog.2018.03.007

Lauffs, M. M., Öğmen, H., & Herzog, M. H. (2017). Unpredictability does not hamper nonretinotopic motion perception. *Journal of Vision*, *17*(9), 6–6. https://doi.org/10.1167/17.9.6

Layton, O. W., Mingolla, E., & Yazdanbakhsh, A. (2012). Dynamic coding of border-ownership in visual cortex. *Journal of Vision*, *12*(13), 8–8.

LeCun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (1988). A theoretical framework for back-propagation. *Proceedings of the 1988 Connectionist Models Summer School*, *1*, 21–28.

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. https://doi.org/10.1016/j.visres.2007.12.009

Levi, D. M., & Carney, T. (2009). Crowding in peripheral vision: Why bigger is better. *Current Biology*, *19*(23), 1988–1993.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. *ArXiv Preprint ArXiv:1901.00945*.

Linsley, D., Kim, J., Ashok, A., & Serre, T. (2020). *RECURRENT NEURAL CIRCUITS FOR CONTOUR DETECTION*. 23.

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision*, *7*(2), 4. https://doi.org/10.1167/7.2.4

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. https://doi.org/10.1167/7.2.24

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

Malania, M., Pawellek, M., Plank, T., & Greenlee, M. W. (2020). Training-Induced Changes in Radial–Tangential Anisotropy of Visual Crowding. *Translational Vision Science & Technology*, *9*(9), 25–25. https://doi.org/10.1167/tvst.9.9.25

Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision*, *15*(8), 16–16. https://doi.org/10.1167/15.8.16

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. https://doi.org/10.1167/16.3.35

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. https://doi.org/10.1167/12.10.13

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision*, *13*(13), 10–10. https://doi.org/10.1167/13.13.10

Manassi, M., & Whitney, D. (2018). Multi-level Crowding and the Paradox of Object Recognition in Clutter. *Current Biology*, *28*(3), R127–R133. https://doi.org/10.1016/j.cub.2017.12.051

Maus, G. W., Fischer, J., & Whitney, D. (2011). Perceived Positions Determine Crowding. *PLOS ONE*, *6*(5), e19796. https://doi.org/10.1371/journal.pone.0019796

Metzger, W. (1936). *Gesetze des Sehens [Laws of seeing].* (Frankfurt am Main, Germany).

Metzger, W., Spillmann, L. T., Lehar, S. T., Stromeyer, M. T., & Wertheimer, M. T. (2006). *Laws of seeing.* Mit Press.

Millin, R., Arman, A. C., Chung, S. T., & Tjan, B. S. (2014). Visual crowding in V1. *Cerebral Cortex*, *24*(12), 3107–3115.

Montaser-Kouhsari, L., & Rajimehr, R. (2005). Subliminal attentional modulation in crowding condition. *Vision Research*, *45*(7), 839–844. https://doi.org/10.1016/j.visres.2004.10.020

Motter, B. C., & Simoni, D. A. (2007). The roles of cortical image separation and size in active visual search performance. *Journal of Vision*, *7*(2), 6–6. https://doi.org/10.1167/7.2.6

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (n.d.). *The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded*. 11.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. (2018). Task-Driven Convolutional Recurrent Models of the Visual System. *ArXiv Preprint ArXiv:1807.00053*.

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience*, *34*(6), 2374–2388.

Neri, P. (2017). Object segmentation controls image reconstruction from natural scenes. *PLoS Biology*, *15*(8), e1002611.

Oberfeld, D., & Stahn, P. (2012). Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. *PloS One*, *7*(10), e48054.

Öğmen, H., & Herzog, M. H. (2010). The Geometry of Visual Perception: Retinotopic and Nonretinotopic Representations in the Human Visual System. *Proceedings of the IEEE*, *98*(3), 479–492. https://doi.org/10.1109/JPROC.2009.2039028

Overvliet, K. E., & Sayim, B. (2016). Perceptual grouping determines haptic contextual modulation. *Vision Research*, *126*, 52–58. https://doi.org/10.1016/j.visres.2015.04.016

Pachai, M. V., Doerig, A. C., & Herzog, M. H. (2016). How best to unify crowding? *Current Biology*, *26*(9), R352–R353. https://doi.org/10.1016/j.cub.2016.03.003

Palmer, S. E. (1992). Common region: A new principle of perceptual grouping. *Cognitive Psychology*, *24*(3), 436–447.

Palmer, S. E., & Beck, D. M. (2007). The repetition discrimination task: An objective method for studying perceptual grouping. *Perception & Psychophysics*, *69*(1), 68–78.

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. https://doi.org/10.1038/89532

Pelli, D. G. (2008). Crowding: A cortical constraint on object recognition. *Current Opinion in Neurobiology*, *18*(4), 445–451.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135. https://doi.org/10.1038/nn.2187

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Petrov, Y., & Meleshkevich, O. (2011). Asymmetries and idiosyncratic hot spots in crowding. *Vision Research*, *51*(10), 1117–1123. https://doi.org/10.1016/j.visres.2011.03.001

Petrov, Y., Popple, A. V., & McKee, S. P. (2007). Crowding and surround suppression: Not to be confused. *Journal of Vision*, *7*(2), 12. https://doi.org/10.1167/7.2.12

Pikler, J. (1917). *Sinnesphysiologische Untersuchungen*. https://books.google.ch/books?id=DlLJeu6AQSsC

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

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

Pollen, D. A. (1999). On the neural correlates of visual perception. *Cerebral Cortex (New York, N.Y.: 1991)*, *9*(1), 4–19. https://doi.org/10.1093/cercor/9.1.4

R Core Team. (2019). *R: A Language and Environment for Statistical Computing* (Vienna, Austria). R Foundation for Statistical Computing. https://www.R-project.org/

Reuther, J., & Chakravarthi, R. (2014). Categorical membership modulates crowding: Evidence from characters. *Journal of Vision*, *14*(6), 5–5. https://doi.org/10.1167/14.6.5

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

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.

Rosenholtz, R., Yu, D., & Keshvari, S. (2019). Challenges to pooling models of crowding: Implications for visual mechanisms. *Journal of Vision*, *19*(7), 15–15. https://doi.org/10.1167/19.7.15

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. https://doi.org/10.1167/8.3.23

Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision*, *9*(2), 5–5. https://doi.org/10.1167/9.2.5

Saarela, T. P., Westheimer, G., & Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *Journal of Vision*, *10*(10), 17–17. https://doi.org/10.1167/10.10.17

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, 3856–3866.

Sayim, B., & Taylor, H. (2019). Letters Lost: Capturing Appearance in Crowded Peripheral Vision Reveals a New Kind of Masking. *Psychological Science*, *30*(7), 1082–1086. https://doi.org/10.1177/0956797619847166

Sayim, B., & Wagemans, J. (2017). Appearance changes and error characteristics in crowding revealed by drawings. *Journal of Vision*, *17*(11), 8–8. https://doi.org/10.1167/17.11.8

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in vernier acuity. *Journal of Vision*, *8*(8), 12–12. https://doi.org/10.1167/8.8.12

Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt Factors Modulate Basic Spatial Vision. *Psychological Science*, *21*(5), 641–644. https://doi.org/10.1177/0956797610368811

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., & Graepel, T. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

Sekuler, A. B., & Bennett, P. J. (2001). Generalized common fate: Grouping by common luminance changes. *Psychological Science*, *12*(6), 437–444.

Silson, E. H., Reynolds, R. C., Kravitz, D. J., & Baker, C. I. (2018). Differential Sampling of Visual Space in Ventral and Dorsal Early Visual Cortex. *The Journal of Neuroscience*, *38*(9), 2294–2303. https://doi.org/10.1523/JNEUROSCI.2717-17.2018

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Solomon, J. A., Felisberti, F. M., & Morgan, M. J. (2004). Crowding and the tilt illusion: Toward a unified account. *Journal of Vision*, *4*(6), 9–9. https://doi.org/10.1167/4.6.9

Strasburger, H. (2005). Unfocussed spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision*, *5*(11), 8–8. https://doi.org/10.1167/5.11.8

Strasburger, H. (2020). Seven Myths on Crowding and Peripheral Vision. *I-Perception*, *11*(3), 2041669520913052. https://doi.org/10.1177/2041669520913052

Strasburger, H., Harvey, L. O., & Rentschler, I. (1991). Contrast thresholds for identification of numeric characters in direct and eccentric view. *Perception & Psychophysics*, *49*(6), 495–508.

Stuart, J. A., & Burian, H. M. (1962). A Study of Separation Difficulty*: Its Relationship to Visual Acuity in Normal and Amblyopic Eyes. *American Journal of Ophthalmology*, *53*(3), 471–477. https://doi.org/10.1016/0002-9394(62)94878-X

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. http://arxiv.org/abs/1312.6199

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient Estimates on Probability Functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787. https://doi.org/10.1121/1.1910407

Ternus, J. (1926). Experimentelle untersuchungen über phänomenale Identität. *Psychologische Forschung*, *7*(1), 81–136.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience*, *14*(5), 869–876. https://doi.org/10.1046/j.0953-816x.2001.01717.x

Thunell, E., van der Zwaag, W., Öğmen, H., Plomp, G., & Herzog, M. H. (2016). Retinotopic encoding of the Ternus-Pikler display reflected in the early visual areas. *Journal of Vision*, *16*(3), 26. https://doi.org/10.1167/16.3.26

Todorović, D. (2007). W. Metzger, Laws of Seeing. *Gestalt Theory*, *29*(2), 176.

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357. https://doi.org/10.1016/0042-6989(92)90227-A

Tootell, R. B. H., Hadjikhani, N. K., Mendola, J. D., Marrett, S., & Dale, A. M. (1998). From retinotopy to recognition: FMRI in human visual cortex. *Trends in Cognitive Sciences*, *2*(5), 174–183. https://doi.org/10.1016/S1364-6613(98)01171-1

Tripathy, S. P., & Cavanagh, P. (2002). The extent of crowding in peripheral vision does not scale with target size. *Vision Research*, *42*(20), 2357–2369. https://doi.org/10.1016/S0042-6989(02)00197-9

Tufte, E. R. (2006). *Beautiful evidence* (Vol. 1). Graphics Press Cheshire, CT.

Vacher, J., & Coen-Cagli, R. (2019). Combining mixture models with linear mixing updates: Multilayer image segmentation and synthesis. *ArXiv Preprint ArXiv:1905.10629*.

Vacher, J., Davila, A., Kohn, A., & Coen-Cagli, R. (2020). Texture Interpolation for Probing Visual Perception. *Advances in Neural Information Processing Systems*, *33*, 22146–22157.

Vacher, J., Mamassian, P., & Coen-Cagli, R. (2019). An Ideal Observer Model for Grouping and Contour Integration in Natural Images. *PERCEPTION*, *48*, 88–88.

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision*, *7*(2), 14–14.

Van Essen, D. C., Lewis, J. W., Drury, H. A., Hadjikhani, N., Tootell, R. B. H., Bakircioglu, M., & Miller, M. I. (2001). Mapping visual cortex in monkeys and humans using surface-based atlases. *Vision Research*, *41*(10), 1359–1378. https://doi.org/10.1016/S0042-6989(01)00045-1

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142.

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. https://doi.org/10.1167/9.2.12

Wagemans, J. (2015). *The Oxford handbook of perceptual organization*. Oxford Library of Psychology.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218.

Wallace, J. M., & Tjan, B. S. (2011). Object crowding. *Journal of Vision*, *11*(6), 19–19. https://doi.org/10.1167/11.6.19

Weymouth, F. W. (1958). Visual sensory units and the minimal angle of resolution. *American Journal of Ophthalmology*, *46*(1), 102–113.

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. https://doi.org/10.1016/j.tics.2011.02.005

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314–1329. https://doi.org/10.3758/BF03194545

Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *Journal of the Optical Society of America A*, *14*(9), 2057. https://doi.org/10.1364/JOSAA.14.002057

World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. (2013). *JAMA*, *310*(20), 2191. https://doi.org/10.1001/jama.2013.281053

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yeh, S.-L., He, S., & Cavanagh, P. (2012). Semantic priming from crowded words. *Psychological Science*, *23*(6), 608–616. https://doi.org/10.1177/0956797611434746

Yeotikar, N. S., Khuu, S. K., Asper, L. J., & Suttle, C. M. (2011). Configuration specificity of crowding in peripheral vision. *Vision Research*, *51*(11), 1239–1248. https://doi.org/10.1016/j.visres.2011.03.016

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. https://doi.org/10.1167/10.10.16

Yeshurun, Y., Rashal, E., & Tkacz-Domb, S. (2015). Temporal crowding and its interplay with spatial crowding. *Journal of Vision*, *15*(3), 11. https://doi.org/10.1167/15.3.11

Yildirim, F. Z., Coates, D. R., & Sayim, B. (2020). Redundancy masking: The loss of repeated items in crowded peripheral vision. *Journal of Vision*, *20*(4), 14–14. https://doi.org/10.1167/jov.20.4.14

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron*, *47*(1), 143–153.

Zhou, H., Friedman, H. S., & Heydt, R. von der. (2000). Coding of Border Ownership in Monkey Visual Cortex. *Journal of Neuroscience*, *20*(17), 6594–6611. https://doi.org/10.1523/JNEUROSCI.20-17-06594.2000

Appendix

Appendix A

Bornet, A., **Choung, O. H.**, Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (2021). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Jornal of vision*, 21(12):10, 1-25.

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

Alban Bornet
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Oh-Hyeon Choung
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Adrien Doerig
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland
Donders Institute for Brain, Cognition and Behaviour,
Nijmegen, Netherlands

David Whitney
Department of Psychology, University of California,
Berkeley, California, USA
Helen Wills Neuroscience Institute,
University of California, Berkeley, California, USA
Vision Science Group, University of California, Berkeley,
California, USA

Michael H. Herzog
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Mauro Manassi
School of Psychology, University of Aberdeen,
King's College, Aberdeen, UK

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 different understanding of crowding, as a global, complex, and multilevel 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.

Citation: Bornet, A., Choung, O.-H., Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (2021). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Journal of Vision*, *21*(12):10, 1–25, https://doi.org/10.1167/jov.21.12.10.

https://doi.org/10.1167/jov.21.12.10 Received February 27, 2021; published November 23, 2021 ISSN 1534-7362 Copyright 2021 The Authors

Downloaded from jov.arvojournals.org on 11/24/2021

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Introduction

In crowding, perception of a target strongly deteriorates when flanking elements are added (Pelli, 2008; Strasburger, Rentschler, & Jüttner, 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, Wilson, & Ellemberg, 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, Lund, Angelucci, Solomon, & Morgan, 2001; Wilkinson et al., 1997), and (3) low-level feature specific (Andriessen & Bouma, 1976; Chung, Levi, & Legge, 2001; Levi, Toet, Tripathy, & Kooi, 1994; Levi, Hariharan et al., 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, Retell, Remington, & Mattingley, 2013; Malania, Herzog, & Westheimer, 2007; Manassi, Sayim, & Herzog, 2012; Vickery, Shim, Chakravarthi, Jiang, & Luedeman, 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; Levne & Sagi, 2010; Malania et al., 2007; Saarela, Westheimer, & Herzog, 2010). Third, crowding was shown to occur at multiple levels along the visual hierarchy (e.g., for objects and faces; Kimchi & Pirkner, 2015; Louie, Bressler, & Whitney, 2007; Sun & Balas, 2015; Xia, Manassi, Nakayama, Zipser, & Whitney, 2020). 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 (for reviews, see: Herzog, Sayim, Chicherov, & Manassi, 2015; Herzog, Sayim, Manassi, & Chicherov, 2016; Herzog & Manassi, 2015; Manassi & Whitney, 2018; see also Banks, Larson, & Prinzmetal, 1979; Banks & White, 1984; Egeth & Santee, 1981; Huckauf, Heller, & Nazir, 1999; Mason, 1982; Mewhort, Marchetti, & Campbell, 1982; Wolford & Chambers, 1983). As a consequence, simple pooling models do not seem adequate to explain this large body of results (Doerig, Bornet, Rosenholtz, Francis, Clarke, & Herzog, 2019; Rosenholtz, Yu, & Keshvari, 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, Balas, & Ilie, 2012), can explain all these effects. In a first stage, the TTM computes V1-like responses from low-level, multiscale, 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 as high-level as previously thought.

Rosenholtz et al. (2019) proposed two ways in which grouping might affect the perception of a crowded stimulus, without requiring explicit visual grouping processes. First, what we call grouping might simply be a collateral effect of high-dimensional pooling. For example, the TTM might "group" together elements that can easily be described using summary statistics. Second, what we call "grouping" might reflect processes that happen after high-dimensional pooling. For example, the high-dimensional pooling stage may reduce the position uncertainty of visual elements (cueing). Moreover, Rosenholtz et al. (2019) proposed that the TTM can also reproduce holistic effects in crowding without requiring high-level feature interactions. The rich information preserved by the high-dimensional pooling stage of the TTM may drive holistic processing (e.g., upright and inverted faces being perceived differently), in a post-perceptual stage.

Here, we tested these hypotheses by probing the TTM behavior on a large body of evidence for high-level effects in crowding (Canas-Bajo & Whitney, 2020; Farzin, Rivera, & Whitney, 2009; Manassi et al., 2012; Manassi, Sayim, & Herzog, 2013; Manassi, Hermens, Francis, & Herzog, 2015; Manassi, Lonchampt, Clarke, & Herzog, 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; Manassi et al., 2013; Manassi, Hermens, Francis, & Herzog, 2015; Manassi, Lonchampt, Clarke, & Herzog, 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, as previously mentioned, Rosenholtz et al. (2019) argued that the grouping effects in crowding (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016) might 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

Downloaded from jov.arvojournals.org on 11/24/2021

no real role in crowding (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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, Nakano, & Rosenholtz, 2009; Rosenholtz et al., 2019).

Stimulus images were taken from (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Canas-Bajo & Whitney, 2020; 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; Manassi et al., 2015), stimuli were displayed on oscilloscopes. Here, we adapted our stimuli to an LCD presentation by having white lines on a black background, as in (Manassi et al., 2013; Manassi et al., 2016).

Model assessment and potential shortcomings

To assess the TTM behavior, following Rosenholtz et al. (2019), we asked participants to perform the original crowding experiments of (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Canas-Bajo & Whitney, 2020; Farzin et al., 2009), but using the mongrels presented in free viewing conditions. All original experiments were two alternative forced choice (2AFC) target discrimination tasks (more detail in the methods subsection of each experiment). To quantify the TTM performance, we measured target discrimination accuracy for each condition. We attempted to address potential shortcomings of our model assessment method in the following ways.

First, we used the code from the official repository to generate the mongrels. The TTM has a variable parameter that needs to be set, namely the radius of the fovea. The code documentation suggests a value between 16 and 32 pixels. The latter value is what was used in Rosenholtz et al. (2019). Because a value of 32 did not yield sufficiently strong crowding in pilot experiments, which would rule out the TTM as a model of crowding, we used a value of 16 pixels. In order to control for ceiling effects, we repeated some experiments with a radius of 32 pixels (details in the subsection of each experiment).

Second, a single mongrel cannot be regarded as the true output of the TTM but merely as an illustration of its behavior. To have a precise measure of the model output, we generated as many mongrels as we could for each stimulus (10 to 200, depending on the number of conditions we needed to run for each experiment). Moreover, we made all generated mongrels available at https://github.com/albornet/TTM_Verniers_Faces_Mongrels.

Third, humans may have strong individual biases in the perception of the mongrels, which may average out existing effects. For this reason, we also used bias-free algorithms to perform the mongrel discrimination tasks (more details in the Methods subsection of each experiment and in the Discussion section).

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 and 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

Downloaded from jov.arvojournals.org on 11/24/2021

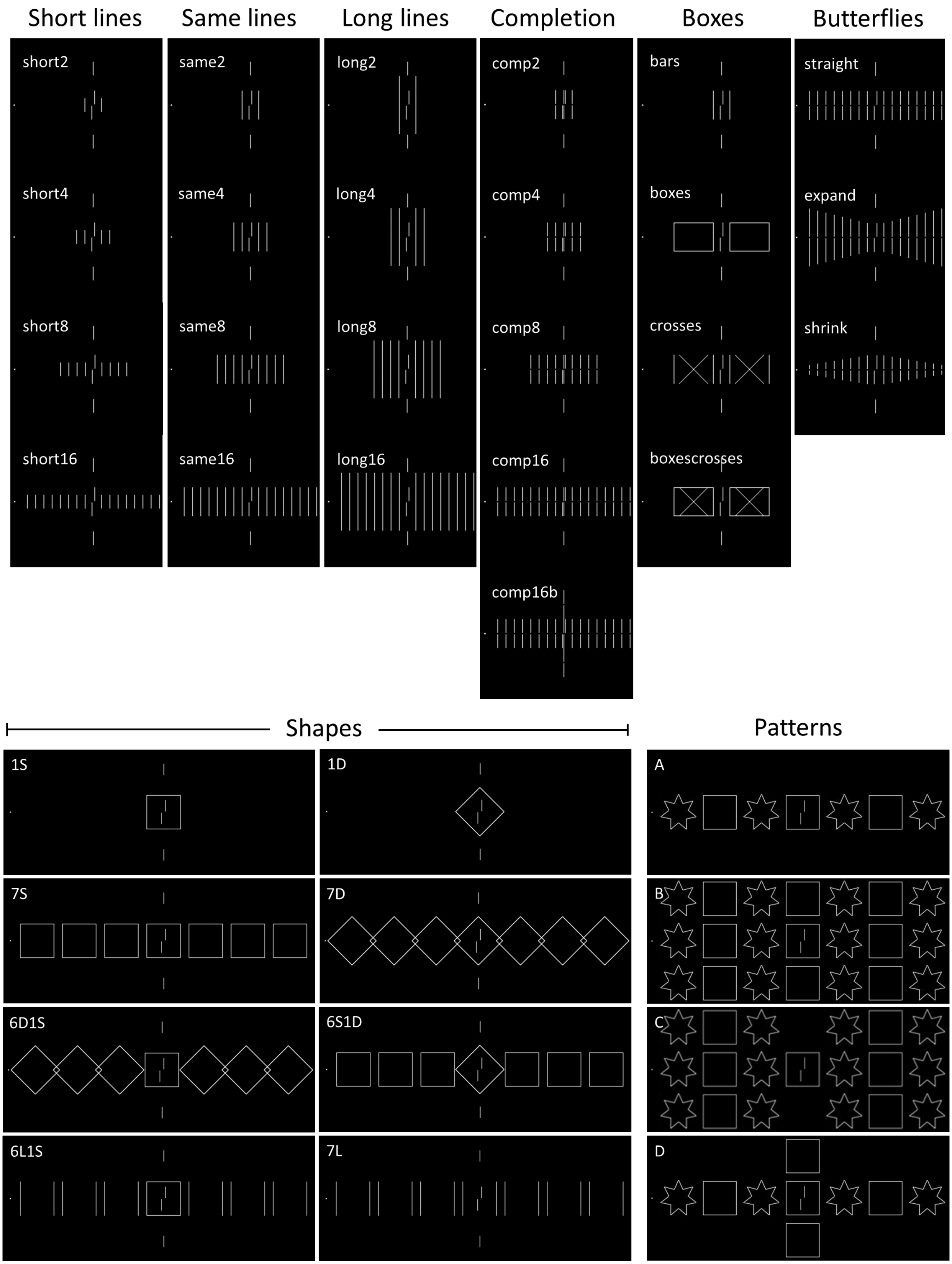


Figure 1. Stimuli used to validate the TTM. In the original experiments, observers were asked to discriminate the offset of a vernier target presented in the right hemifield and in the periphery (here shown in the center of each image), while looking at a fixation dot.

→

Downloaded from jov.arvojournals.org on 11/24/2021

←

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 degrees in the "Lines," "Boxes," "Completion," and "Butterflies" experiments, and 9 degrees in the "Shapes" and "Patterns" experiments. Note that, in all original experiments except "Patterns", two vertical lines (pointers) were added above and below the vernier target to reduce target location uncertainty.

---

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).

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 degrees 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 degrees and a vernier offset size of 14 arcmin. These vernier offsets correspond to approximately five times the thresholds measured in the original experiments for the unflanked conditions (vernier alone).

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; Manassi et al., 2013; Manassi et al., 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 labeled.

Thirteen observers performed this task (6 men and 7 women, 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 as the sum of the point-wise multiplication between the template and the part of the mongrel image that lay under the target template centered at that location. 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 (Equation 1).

$$M^s(i,j) = \mathrm{e}^{-(D(i,j)/\sigma)^2} \cdot \sum\nolimits_{k,l} T^s_{k,l} \cdot I_{i+k,j+l} \qquad (1)$$

$M^s(i,j)$ was the weighted match value of the s-sided vernier template at location (i,j), $T^s_{k,l}$ 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.

Downloaded from jov.arvojournals.org on 11/24/2021

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 (see Figure 2, left). Hence, adding flankers can lead to nonmonotonic 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 (see Figure 2, center and right). The TTM performance differs from human data, in which adding flankers reduced crowding strength in certain conditions.

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 (see Figure 3, left, straight versus comp16). Crowding was strong for control conditions in which a longer mask was used or using a same-length mask but having only two vernier flankers (see Figure 3, left, comp16b and 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 (see Figure 3, center and right, straight versus comp16). However, crowding strength also decreased when using a longer mask or having only two vernier flankers (see Figure 3, center and right, straight versus comp16b and comp2), and gradually increased when adding more flankers (Supplementary Information Figure SA), showing that the configuration played no role.

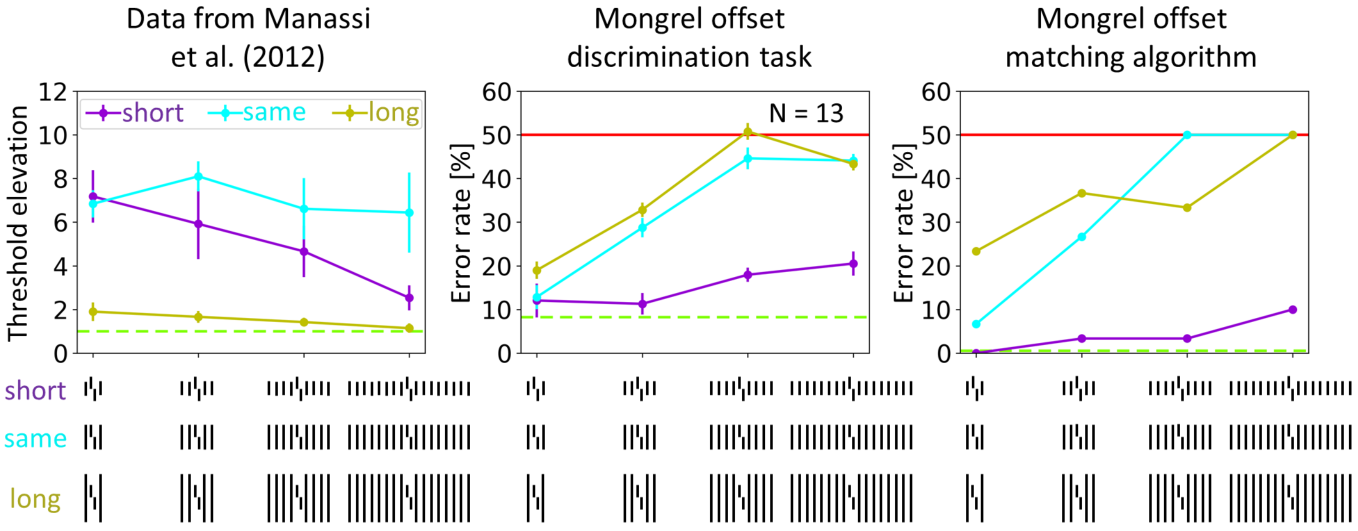
Boxes and crosses experiment

In Manassi et al. (2012), crowding was strong when the vernier target was flanked by two same-length flankers (see Figure 4, left, bars). Crowding decreased when adding flankers to form boxes or boxes containing a cross (see Figure 4, left, boxes and boxes and crosses), but stayed high when the added flankers were not embedded in box shapes (see Figure 4 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 (see Figure 4, center and right, bars versus boxes and crosses and boxes and crosses), regardless of the configurations.

Shapes experiment

In Manassi et al. (2013), crowding was strong when the vernier target was flanked by a single square (see


Data from Manassi et al. (2012)

Mongrel offset discrimination task

Mongrel offset matching algorithm

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.

Downloaded from jov.arvojournals.org on 11/24/2021

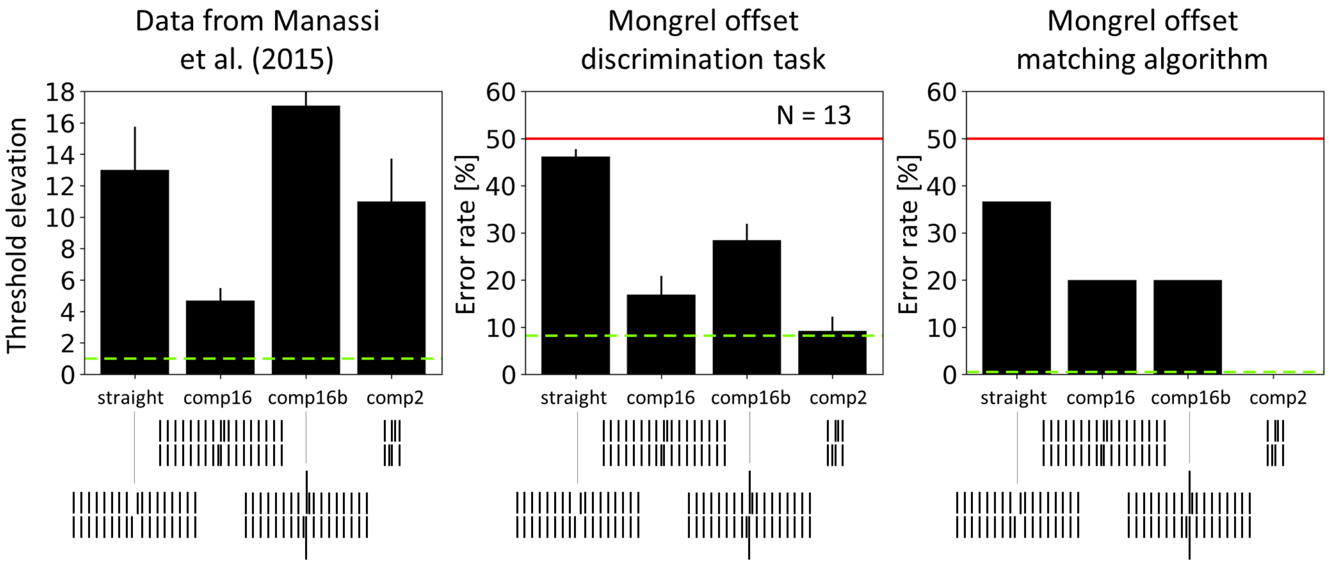


Figure 3. Completion. Left. Data from Manassi et al. (2015). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. 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.

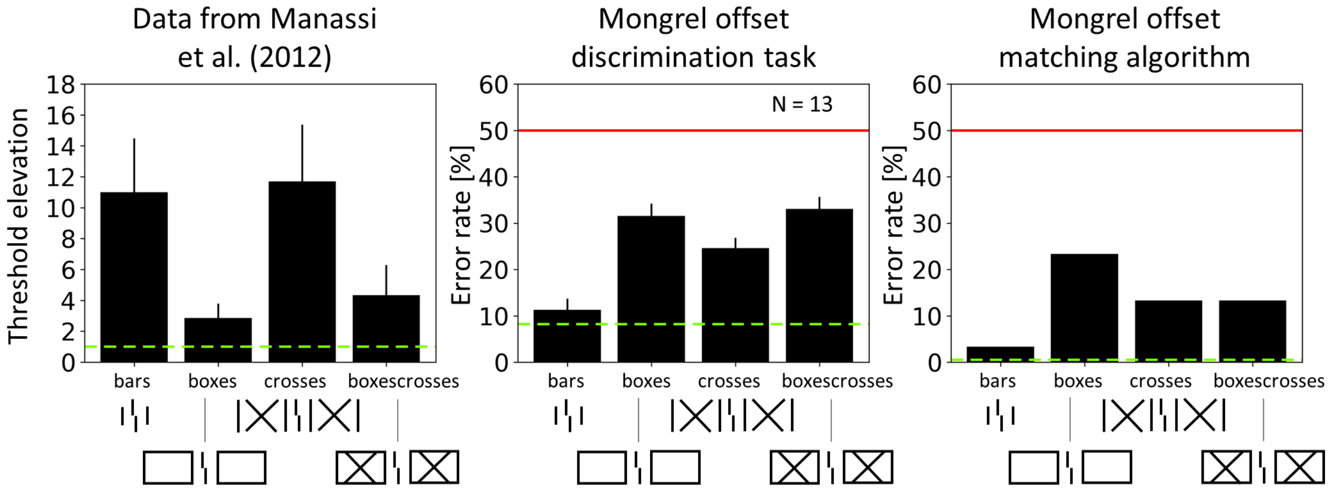


Figure 4. Boxes and crosses. Left. Data from Manassi et al. (2012). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. 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 that the y-axis labels are different.

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 (see Figure 5, left, 7S versus 7D1S). Crowding was strong in control conditions (see Figure 5, left, 7L and 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 (see Figure 5, center and right). A similar pattern was found using diamonds instead of squares (Supplementary Information Figure SB).

Downloaded from jov.arvojournals.org on 11/24/2021

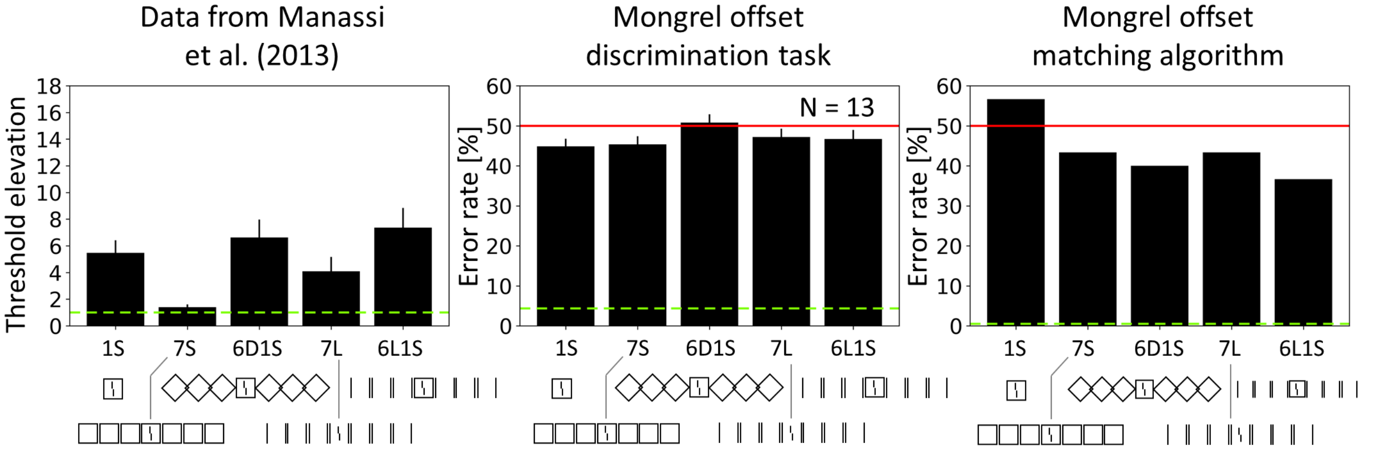


Figure 5. Shapes. Left. Data from Manassi et al. (2013). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 9 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. 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.

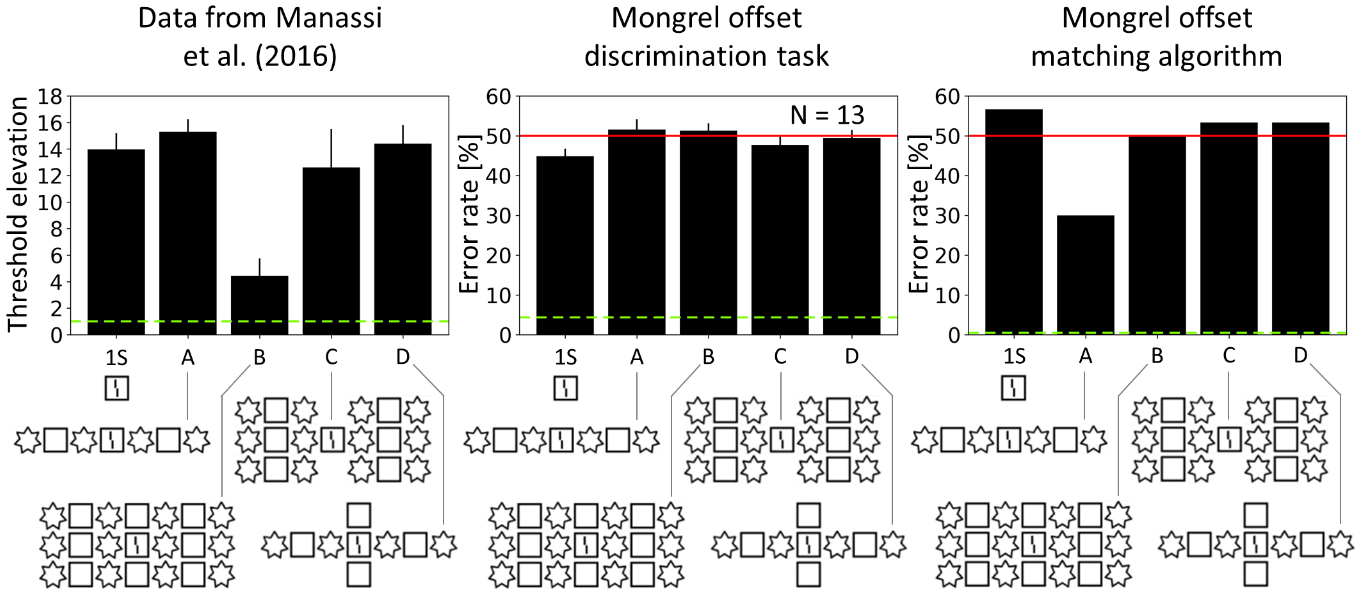


Figure 6. Patterns. Left. Data from Manassi et al. (2016). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 9 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. 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.

Pattern experiment

In Manassi et al. (2016), crowding was strong when the vernier was embedded in a single square (see 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 (see Figure 6, left, A versus B). Crowding was strong in both control conditions (see Figure 6, left, C and 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 (see Figure 6, center and right).

Downloaded from jov.arvojournals.org on 11/24/2021

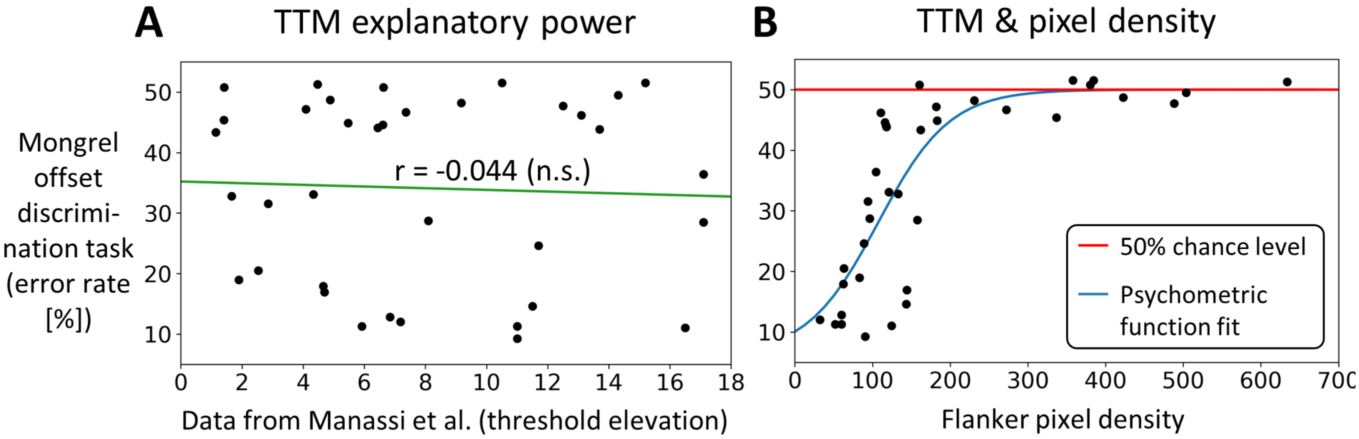


Figure 7. (**A**) TTM performance in the mongrel vernier offset discrimination task showed no correlation ($r = -0.044$, $p = 0.799$, $BF_{01} = 4.672$; Ly, Verhagen, & Wagenmakers, 2016; Rouder, Speckman, Sun, Morey, & Iverson, 2009) with the original data from (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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 (Supplementary Information Figure SF). Fitting the data with a psychometric function (see Equation 3 in Supplementary Information SL), we found a strong correlation between the TTM and the fitted performance ($r(34) = 0.796$, $p < 0.001$, $BF_{10} > 10^6$).

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 and 6, as well as Supplementary Information Figure SB). 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 (Supplementary Information Figure SC).

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

TTM and prediction power

As a global measure of the explanatory power of the TTM for each condition of (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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 value $= 0.799$), indicating that none of the reported results can be explained by the TTM. A similar correlation was found using the template matching algorithm (Supplementary Information Figure SE).

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 Supplementary Information SL). The error rate increased with flanker density (see Figure 7B). Fitting the data with a psychometric function (see Equation 3 in Supplementary Information SL), 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.

Downloaded from jov.arvojournals.org on 11/24/2021

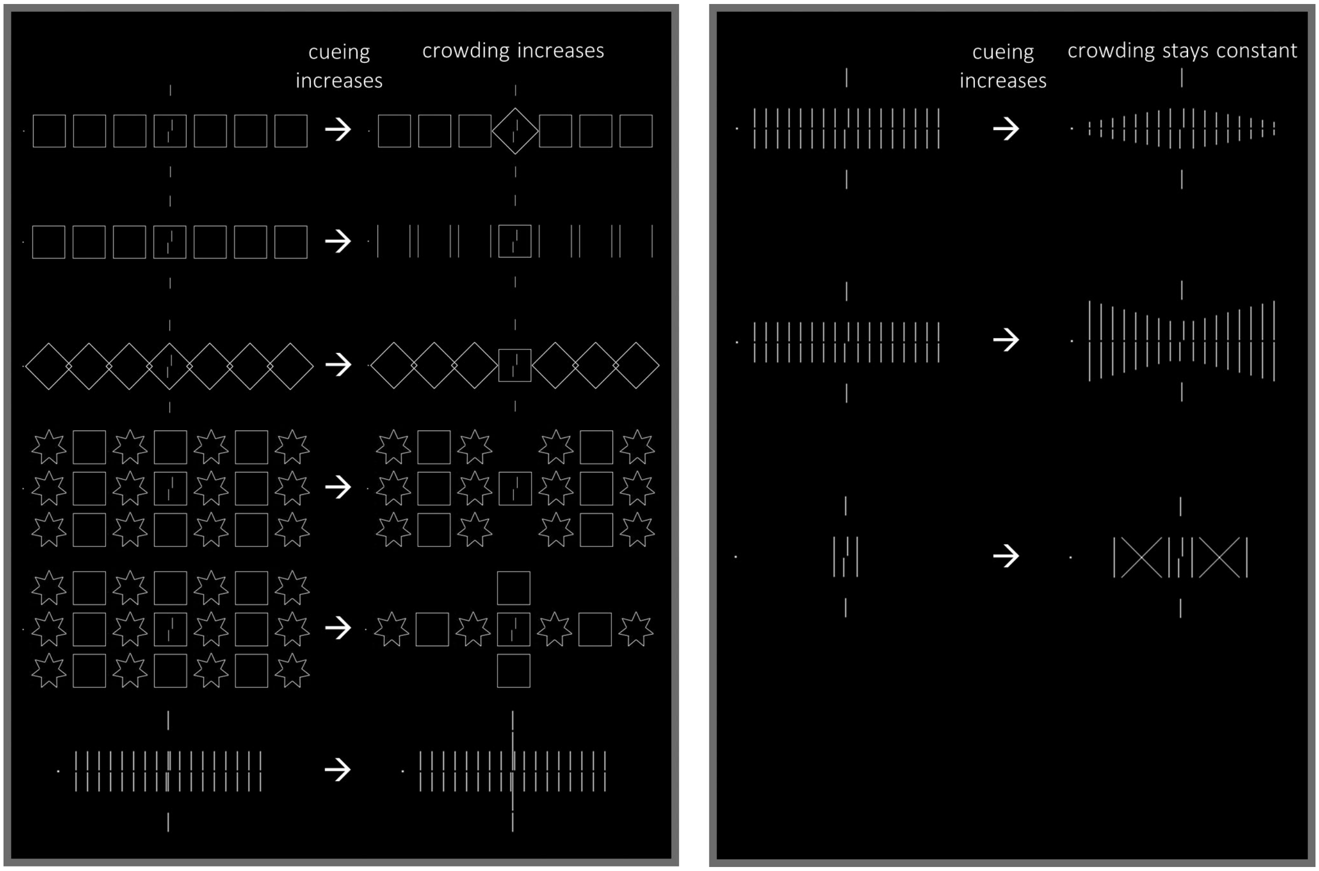


Figure 8. Right column, for both panels. Conditions in which the target location is weakly cued by the flanker configuration. Left Column, for both panels. Conditions in which the target location is strongly cued by the flanker configuration. If cueing had a strong impact on target discrimination performance, 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016).

Grouping effects and target cueing

Rosenholtz et al. (2019) argued that the results in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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 right side than on the left side. According to the cueing argument, crowding should be weaker on the right side compared to the left 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).

Second, in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 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; Manassi et al., 2013; Manassi et al., 2015), the aim was to minimize the target location uncertainty. Rosenholtz et al.

Downloaded from jov.arvojournals.org on 11/24/2021

(2019) argued that these pointers may instead increase crowding by creating multiple offsets among vernier, flankers and pointers lines (see figure 17 in Rosenholtz et al., 2019). However, in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015), the pointers are actually quite far from the vernier, making this offset confusion argument unlikely (see Supplementary Information Figure SG). 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 (Supplementary Information Figure SH).

Third, the effects measured in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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, Kohnen, Barton, & Awh, 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, Westheimer, & Herzog, 2008; Sayim, Westheimer, & Herzog, 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, Chin, & Chung, 2013; Coates, Levi, Touch, & Sabesan, 2018; Danilova & Bondarko, 2007; Flom, Heath, & Takahashi, 1963; Lev, Yehezkel, & Polat, 2014; Lev & Polat, 2015; Sayim, Greenwood, & Cavanagh, 2014; Siderov, Waugh, & Bedell, 2013; Westheimer & Hauske, 1975; but see Levi, Hariharan et al., 2002; Levi, Klein, & Hariharan, 2002), as well as grouping processes acting in foveal (Banks & White, 1984; Bock, Monk, & Hulme, 1993; Tannazzo, Kurylo, & Bukhari, 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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, Sayim, Westheimer, & Herzog, 2009).

TTM and face crowding

In the previous section, we showed that the TTM cannot explain the grouping effects found in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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 (Figure 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, Wang, Shao, Zhang, & He, 2020; Grützner, Uhlhaas, Genc, Kohler, Singer, & Wibral, 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, Apthorp, Aagten-Murphy, & Alais, 2011): upright faces are recognized more easily than inverted ones (Farah, Tanaka, & Drain, 1995; Kanwisher, Tong, & Nakayama, 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, Melloni, & Schurger, 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

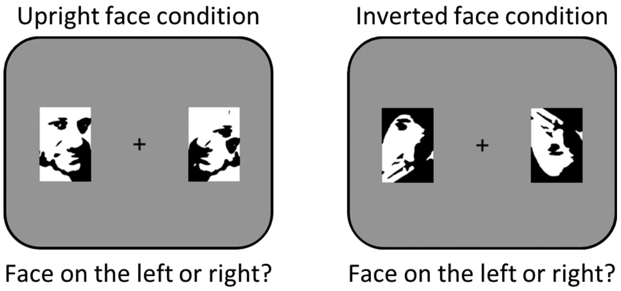


Figure 9. **Single face discrimination task.** Observers were asked to discriminate which of the two images was a face (left or right, 2AFC), by pressing the left or right arrow, while fixating the central cross. Across the experiment, the face could be either upright or inverted. In these examples, an upright face is presented on the left side (left panel), and an inverted face is presented on the right side (right panel). Mooney faces reprinted from Schwiedrzik et al. (2018). Distributed under a CC-BY license.

Downloaded from jov.arvojournals.org on 11/24/2021

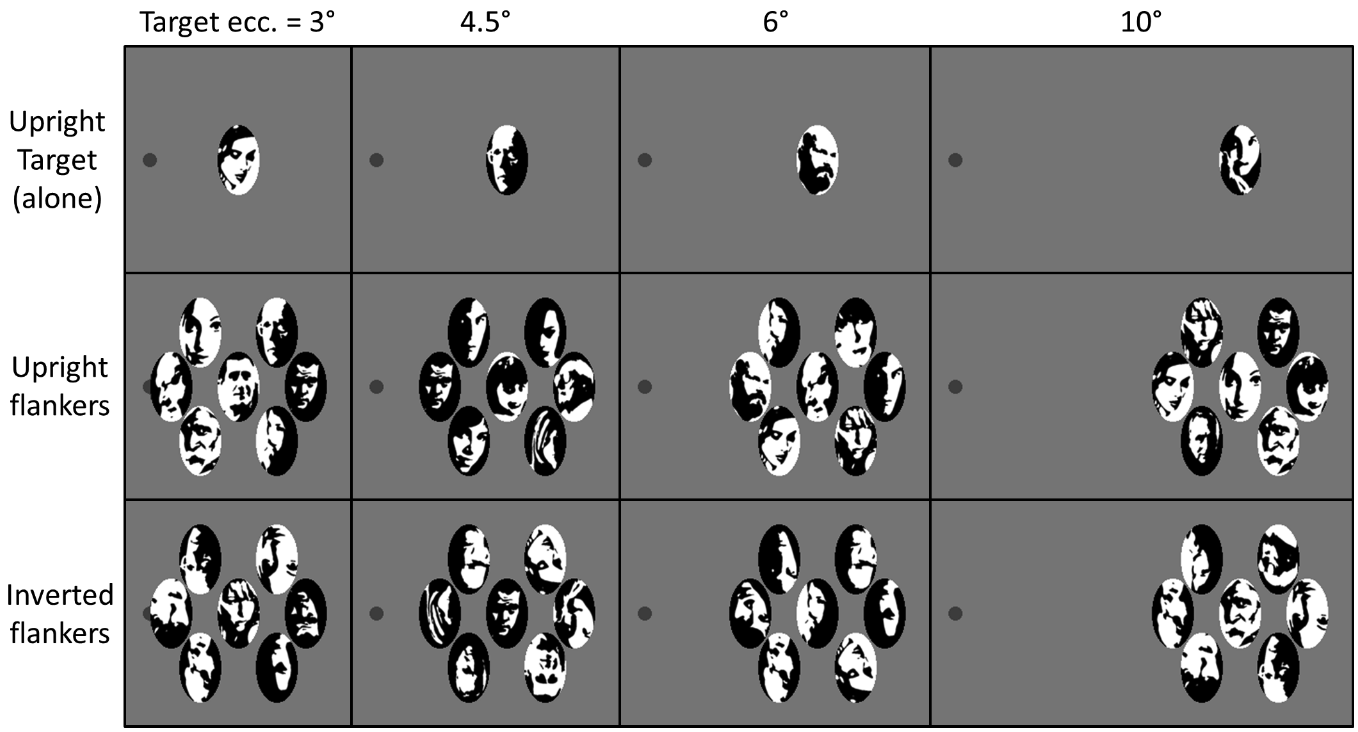


Figure 10. Examples of stimuli used in the face crowding task. There were three main conditions (upright target alone, target with upright flankers or target inverted flankers) presented at four different eccentricities. Mooney faces reprinted from Schwiedrzik et al. (2018). Distributed under a CC-BY license.

of visual processing (inversion effect in a single face discrimination task; Bayle, Schoendorff, Hénaff, & Krolak-Salmon, 2011; Boucart, Lenoble, Quettelart, Szaffarczyk, Despretz, & Thorpe, 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 (see Figure 9). Both images subtended a visual angle of 6 degrees by 4.2 degrees and were presented at the same eccentricity on both sides (6 degrees, 10 degrees, 14 degrees, or 18 degrees). One image was always a Mooney face, whereas the other one was always a scrambled version of the same face. Mooney faces were taken from Schwiedrzik et al. (2018), with permission (freely available at https://doi.org/10.6084/m9.figshare.5783037). 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, as unlimited viewing time has no effect on crowding (Wallace, Chiu, Nandy, & Tjan, 2013). The distance to the screen was 64 cm.

There were five different faces, for a total of 20 different stimuli per eccentricity (2 sides, 2 face orientations, and 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 versus inverted face) and eccentricity, we computed discrimination performance (error rate = 1-accuracy) and the corresponding standard error of the mean, computed over human observers.

In order to validate the TTM, we tested mongrel images with the same single face discrimination task as in Canas-Bajo and 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

Downloaded from jov.arvojournals.org on 11/24/2021

of 200 mongrels shown per eccentricity. Seven observers (2 men and 5 women, 25.4 ± 1.2 years old) performed the task. For each condition (upright versus 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.

In addition, we measured the TTM performance for each condition with a template matching algorithm (Supplementary Information Figure SJ). As for the Vernier offset matching algorithm, a face target was searched in the mongrels by sliding target face templates over the image. The algorithm answered either left or right, as the side of the image on which the best matching score was obtained over all possible target face templates (see Equation 1 for the detailed computation). Accuracy was defined as the percentage of correct answers.

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), with permission (freely available at https://doi.org/10.6084/m9.figshare.5783037). The size of the faces was the same as in Farzin et al. (2009; i.e., 1.53 degrees by 2.48 degrees). 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, and inverted flankers) and four different target eccentricities (3 degrees, 4.5 degrees, 6 degrees, and 10 degrees). Compared to the original experiment, we had an additional eccentricity (4.5 degrees) 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 × 3 flanking conditions x 4 eccentricities). Ten different mongrels were generated for each stimulus, for a total of 2400 unique samples shown to every participant. Seven observers (2 men and 5 women, 25.4 ± 1.2 years old) performed the task.

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 behavioral data of Farzin et al. (2009).

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 Equation 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; Supplementary Information Figure SK).

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 only 0.4% (full model: $\mathrm{r_m}^2 = 0.243$ and $\mathrm{r_c}^2 = 0.696$ and the reduced model: $\mathrm{r_m}^2 = 0.239$ and $\mathrm{r_c}^2 = 0.692$).

Observers were able to discriminate an upright/inverted face from a scrambled face at all tested eccentricities (see Figure 11A). Crucially, observers' accuracy was higher for upright than inverted faces (see Figure 11A, upright versus inverted), indicating a differential processing of inverted (low-level) and upright (holistic) faces, even at 18 degrees of 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 (see 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 section for details). Data were analyzed using a linear mixed effect

Downloaded from jov.arvojournals.org on 11/24/2021

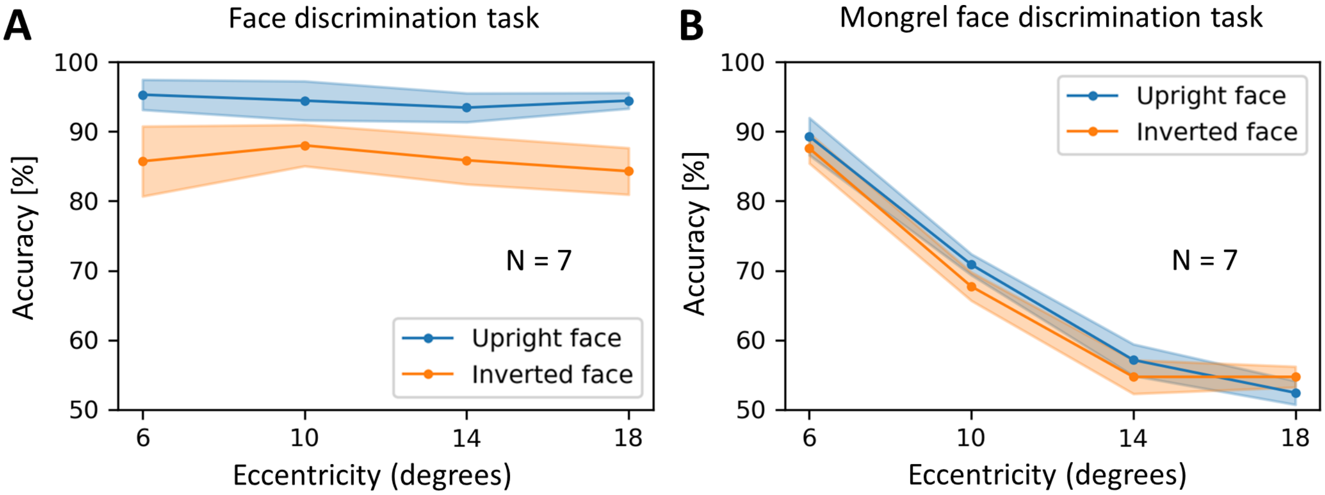


Figure 11. TTM and single Mooney face discrimination. (**A**) Face discrimination task. Observers were asked to discriminate an upright/inverted face from a scrambled face at all tested eccentricities. Accuracy remained on a constant high level for all eccentricities. Crucially, accuracy was higher for 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.

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: $r_m^2 = 0.798$ and $r_c^2 = 0.802$ and the reduced model: $r_m^2 = 0.796$ and $r_c^2 = 0.800$).

These results show that the face discrimination performance in the TTM decreased with increasing eccentricity, contrary to the behavioral results (see Figure 11, A versus B). More importantly, there was no difference between the upright and inverted mongrel face conditions (see Figure 11B, orange versus blue). The lack of inversion effect shows that the high-dimensional pooling stage of the TTM does not preserve rich enough information to support holistic processing in a later post-perceptual stage, as suggested by Rosenholtz et al. (2019).

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 (Supplementary Information Figure SI). Moreover, we also quantified the TTM performance using a template matching algorithm and obtained qualitatively similar results (Supplementary Information Figure SJ).

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 (see 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 (see 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 (see 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

Downloaded from jov.arvojournals.org on 11/24/2021

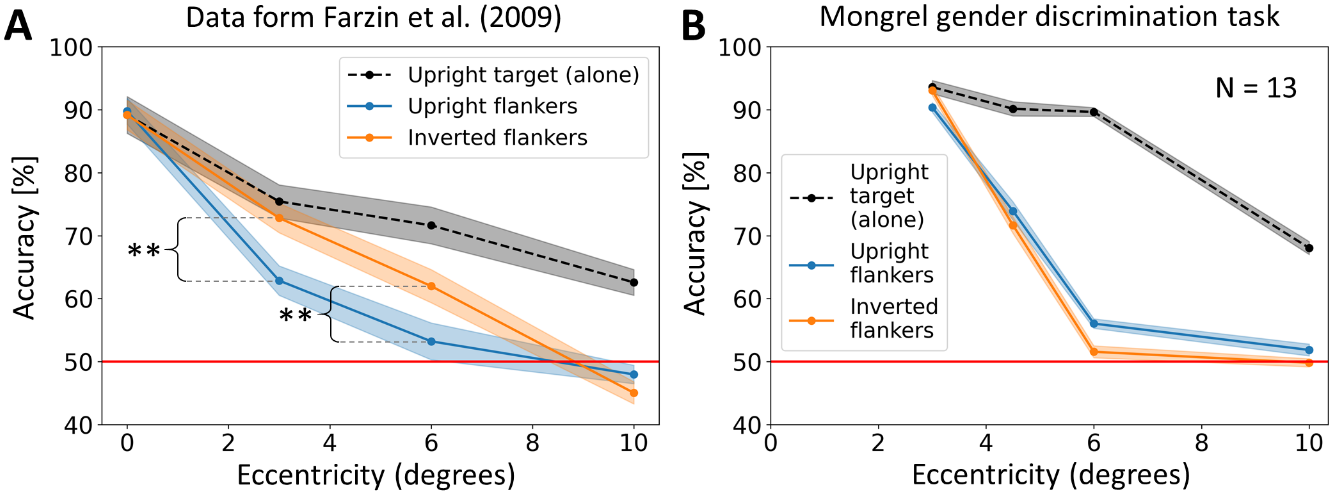


Figure 12. TTM and crowding of Mooney faces. (**A**) Face crowding task, data from Farzin et al. (2009). Target discrimination performance decreased when eccentricity increased. When the target face was flanked by inverted faces, crowding increased with increasing eccentricity (orange). When the target 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 analyzing 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.

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 degrees and 6 degrees 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 section for details, gender face discrimination task). The results (see Figure 12B) were analyzed using a linear mixed effect model, with eccentricity and face orientation (upright versus 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: $r_m^2 = 0.691$ and $r_c^2 = 0.691$ and the reduced model: $r_m^2 = 0.689$ and $r_c^2 = 0.689$).

As in Farzin et al. (2009; see Figure 12A), TTM performance decreased with eccentricity (see 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. In addition, when TTM does predict a trending difference, it is often in a direction opposite that in the empirical data (blue-above-orange in Figure 12B compared to orange-above-blue in Figure 12A). We also quantified the TTM performance using a template matching algorithm and obtained qualitatively similar results (Supplementary Information Figure SK). 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 Figures 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 (see Figure 11) and multiple faces (see 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 (see Figure 7). In this light, the TTM may fail to explain a broad array of findings in the peripheral face recognition literature (Boucart et al., 2016; Farzin et al., 2009; Kovács, Knakker, Hermann, Kovács, & Vidnyánszky, 2017; Kreichman, Bonneh, & Gilaie-Dotan, 2020).

Downloaded from jov.arvojournals.org on 11/24/2021

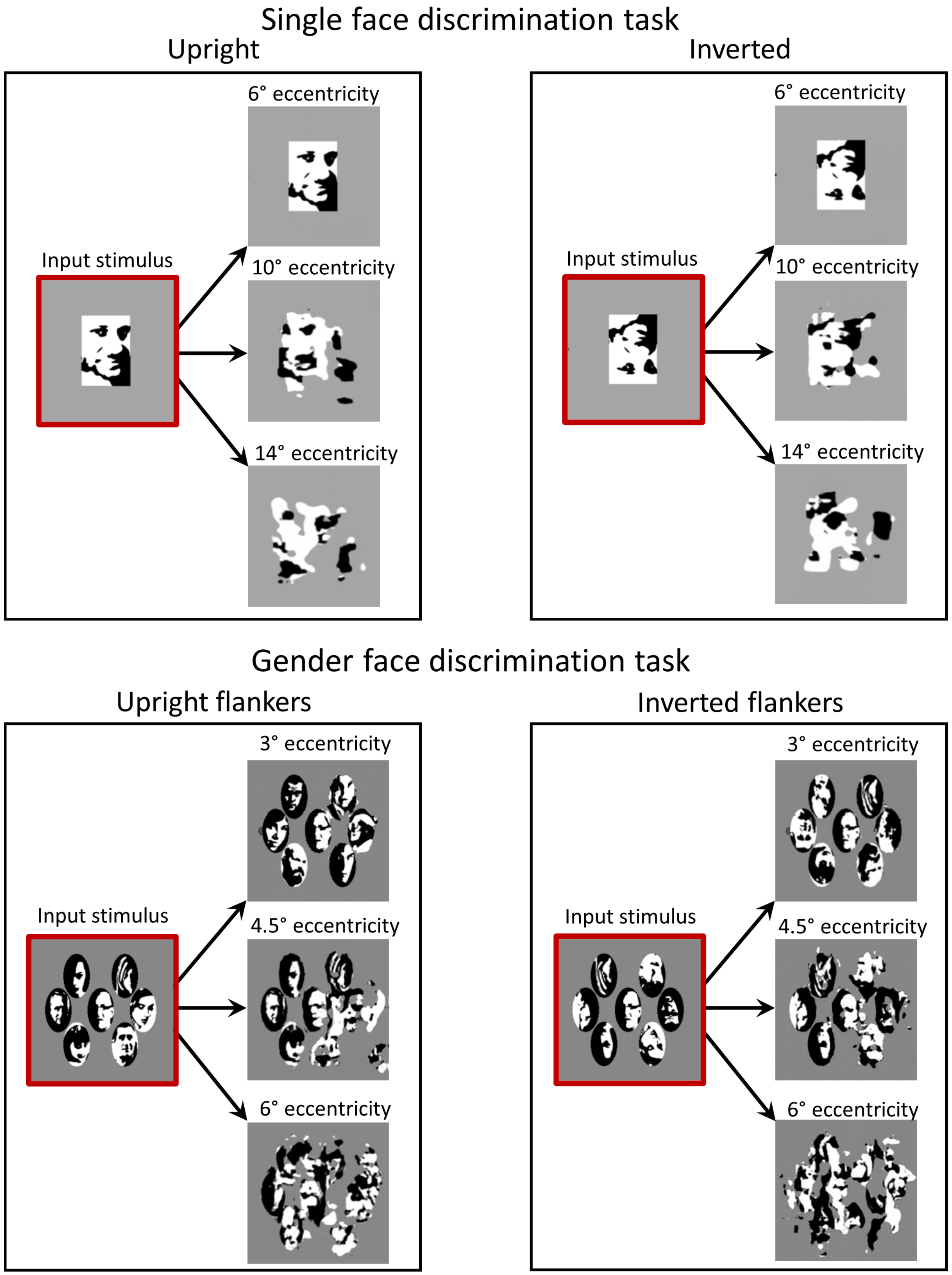


Figure 13. **TTM mongrel examples used in the single face and gender face discrimination tasks.** The stimuli (TTM input) are highlighted in red. To give a representative sample of the TTM outputs for each example, we show mongrels for different eccentricities. Note that we cropped the mongrels for ease of comparison. All mongrels can be found at https://github.com/albornet/TTM_Verniers_Faces_Mongrels. Mooney faces reprinted from Schwiedrzik et al. (2018). Distributed under a CC-BY license.

Downloaded from jov.arvojournals.org on 11/24/2021

Discussion

Classic models describe crowding as a local and low-level phenomenon (Greenwood, Bex, & Dakin, 2009; Levi, Hariharan et al., 2002; Nandy & Tjan, 2012; Parkes et al., 2001; Van den Berg, Roerdink, & Cornelissen, 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Manassi & Whitney, 2018; Saarela et al., 2009; Saarela, et al., 2010; Whitney & Levi, 2011; for older studies about high-level effects in crowding, see also Banks et al., 1979; Banks & White, 1984; Egeth & Santee, 1981; Huckauf et al., 1999; Mason, 1982; Mewhort et al., 1982; Wolford & Chambers, 1983). More recently, it was shown that crowding is affected by emotional conditioning of the flankers (Pittino, Eberhardt, Kurz, & Huckauf, 2019) or by the high-level semantic information of visual scenes (Gong, Xuan, Smart, & Olzak, 2018). This large body of evidence for high-level effects in crowding suggest that current models of vision need to be radically updated. However, against this 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.

Importantly, the current work is not about the TTM only. Instead, it asks the question whether a sophisticated pooling stage can preserve rich enough information about the stimulus to drive the global aspects of crowding in a post-perceptual stage. This argument has implications that go beyond a simple model controversy. For example, global configuration does not need to affect low-level information if Rosenholtz et al. (2019) is correct. In the following, we describe implications from our two sets of data on grouping effects and face recognition.

TTM and 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), in which: (1) increasing the number of flankers sometimes reduces crowding strength (see Figure 2); (2) adding a single element has a dramatic effect on crowding strength (see Figure 3; completion effect); (3) the overall configuration of the flankers determines crowding (see Figure 4); (4) high-level processing strongly affects low-level processing (see Figure 5), and (5) adding flankers beyond Bouma's window strongly modulates crowding strength (see 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; Saarela et al., 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, Sayim, Manassi, & Herzog, 2020). In the TTM, crowding strength was never reduced, when additional flankers were added, regardless of flanker configuration (see Figures 2 to 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 (see Figure 3, center and right, straight versus 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 (see Figure 3, center and right, comp16b and 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 (see Supplementary Information Figure SA), 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), we found no significant correlation (see Figure 7A). Moreover, we found that the TTM performance strongly correlates with the amount of flankers around the target (see Figure 7B), similar to a simple pooling model. Of course, this does

Downloaded from jov.arvojournals.org on 11/24/2021

not mean that the TTM is only measuring flanker pixel density, but rather that this factor is crucial in driving crowding strength and stimulus appearance in the TTM. Still, it seems that the TTM is blind to complex configurations and grouping cues. We propose that the main reason for this lies in the model architecture (i.e., feedforward pooling cannot explain high-level effects in crowding; Bornet, Doerig, Herzog, Francis, & Van der Burg, 2021; Doerig, Bornet, Choung, & Herzog, 2020; Doerig et al., 2019; Doerig, Schmittwilken et al., 2020; Choung, Bornet, Doerig, & Herzog, 2021).

There are several other 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 expressed by Wallis, Funke, Ecker, Gatys, Wichmann, and Bethge (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 (see 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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, because effects of contextual grouping were also found in foveal vision (Saarela & Herzog, 2008; Sayim et al., 2010; Sayim, Westheimer, & Herzog, 2011; Sayim, Manassi, & Herzog, 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; Coates et al., 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 and face crowding

In another set of experiments (see 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; see Figure 11A, upright versus 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., see Figure 11B), thus becoming more unrecognizable when increasing in eccentricity (Martelli, Majaj, & Pelli, 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 would allow to discriminate a face from a non-face object is irretrievably lost in the pooling stage of the TTM. 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 (see Figure 11B). This is in contradiction with the results of the single face discrimination task we performed (see Figure 11A; Canas-Bajo & Whitney (2020), and with recent evidence that stimulus information on several levels of visual processing can survive crowding and influence subsequent perceptual judgments (Faivre & Kouider, 2011a; Faivre & Kouider, 2011b), including face-level information (Kouider, Berthet, & Faivre, 2011).

Next, we focused on holistic face crowding (as found in Experiment 6 of Farzin et al., 2009; see 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.

Downloaded from jov.arvojournals.org on 11/24/2021

Rosenholtz et al. (2019) suggested that the TTM could predict these results without requiring high-level feature interactions. Instead, holistic effects might be driven, in a post-perceptual stage, by the rich information that survives high-dimensional pooling in the TTM.

We tested this hypothesis in practice. Using a mongrel gender crowding discrimination task (see Figure 10), we showed that the TTM did not reproduce holistic face crowding (see Figure 12B). Although crowding occurred in the TTM when face flankers were added, there was no effect of flanker face orientation on the TTM performance. In other words, the high-dimensional pooling stage of the TTM did not preserve enough information to drive holistic processing in a post-perceptual stage. This result gives more support to the hypothesis that crowding 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; Rosenholtz et al., 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 (see 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. 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. Indeed, if participants can avoid inadvertently reporting the gender of an inverted flanker face if it is swapped for the target due to location uncertainty, it means that this face needs to be identified as an inverted face. This requires holistic processing, especially for Mooney faces (which cannot be identified using low-level cues). Moreover, the results we obtain in the gender discrimination task (see Figure 12B) suggest that this is not what happens in the TTM.

Model assessment method

It could 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, we used only parameters in the range preconized by the code documentation of the TTM (fovea radius with any value between 16 and 32 pixels). We originally used a fovea radius value of 32 pixels, which is what was used in Rosenholtz et al. (2019). However, for many of the tested conditions (especially with few flankers), the mongrels were almost untouched, which would have led to 100% accuracy, merely invalidating the TTM (i.e., no crowding). In addition, it may have obscured complex model behaviors, because of ceiling effects. For this reason, we decreased the fovea radius parameter from 32 to 16 pixels to increase crowding in all conditions (the main results reported in the current work). Still, for most stimuli that included large flanker configurations at large eccentricities (Shapes and Patterns experiments; see Figures 5 and 6, as well as Supplementary Information Figure SB), 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 (see Supplementary Information Figure SC).

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, Huang, Yigit-Elliott, & Rosenholtz, 2015). Nevertheless, to control for unwanted human biases, we also quantified performance using a template matching algorithm (see Methods section for details). This did not change the results qualitatively (see Figures 2 to 6, as well as Supplementary Information Figures SA, SB, SC, SD, SE, SJ, and SK). The measured performances were similar to what was measured in the behavioral tasks, and none of the high-level effect of crowding were reproduced.

We want to point out that the template matching algorithms do not aim at reproducing human behavior results. They are an alternative (and more objective) way to measure TTM behavior, and to probe the information present in the model after high-dimensional pooling. Ultimately, the goal is to understand human perception, hence the main results of the present work are the ones that come from the human mongrel discrimination tasks. Nevertheless, we still tried to give the TTM an extra chance to reproduce uncrowding or holistic effects that could have been obscured by individual differences or biases of the participants during the mongrel discrimination tasks.

Downloaded from jov.arvojournals.org on 11/24/2021

Model improvements

The TTM could account for a variety of perceptual properties of human vision, such as visual search (Alexander, Schmidt, & Zelinsky, 2014; Chang & Rosenholtz, 2016; Rosenholtz, 2011; Rosenholtz, Huang, Raj et al., 2012), gist perception and change blindness (Rosenholtz, 2014; Rosenholtz, et al., 2016; Ehinger & Rosenholtz, 2016), or visual metamers (Freeman & Simoncelli, 2011). Moreover, simply by using a rich set of image statistics, the TTM can explain many properties of visual crowding, such as substitution effects, its relationship to feature binding, or the selectivity of illusory feature conjunction (Keshvari & Rosenholtz, 2016). Finally, other models like the TTM that are based on image statistics can explain results from a large range of stimuli and tasks (Heeger & Bergen, 1995; Malik & Perona, 1990; Portilla & Simoncelli, 2000; Zhang et al., 2015; Ziemba & Simoncelli, 2021). Hence, our results should not be taken as a complete invalidation of the TTM or related image-statistic based models. Rather, they suggest that, to fully capture human behavior, 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; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 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, Kaiser, Kroner, Falotico, Ambrosano, Cantero, Herzog, & Francis, 2019; Bornet et al., 2021; Doerig, Bornet et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken et al., 2020; Francis, Manassi, & Herzog, 2017; Wallis, Funke, Ecker, Gatys, Wichmann, & Bethge, 2019). Along the same lines, it was shown that perceptual grouping is crucial to understand contextual effects in naturalistic scenes (Herrera-Esposito, Coen-Cagli, Gomez-Sena, 2021). Again, summary-statistics models (Balas et al., 2009; Freeman & Simoncelli, 2011; Parkes et al., 2001; Rosenholtz, 2016; Rosenholtz et al., 2019) could not predict this body of results (Herrera-Esposito et al., 2021).

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 include high-level statistics in high-dimensional pooling models, such as the TTM. Depending on the stimulus, interaction might occur at different levels of the visual processing hierarchy.

Alternatively, Chaney, Fischer, and Whitney (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.

We would like to point out that one of the advantages of the TTM is that it can easily be tested on various paradigms. The model provides a direct visualization of its output, which is not the case for most proposed models of vision. Importantly, the TTM does not need to be adapted or re-trained for new stimuli. This contrasts with, for example, the capsules network of Doerig, Schmittwilken et al. (2020), which needs to re-learn how to group stimuli for any new paradigm. This strong point of the TTM is also why it is easier to falsify it, as for example in the present work.

Finally, we cannot rule out that in the future, more complex or more flexible statistics may be used in the TTM to show that the model can exhibit uncrowding or holistic processing. For example, deep neural networks trained on natural images may be used as a source of complex summary statistics relevant to human perception (Ziemba & Simoncelli, 2021). However, we have reasons to believe that this will not be the case. Indeed, pooling is by nature ill-suited for this task, because adding more flankers always increases interference with the target representation. We do not see how this hurdle can be overcome. For example, feedforward convolutional neural networks, who are explicitly optimized for image recognition in a pooling framework, are biased towards local features (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos, Rubisch, Michaelis, Bethge, Wichmann, & Brendel, 2018; Wallis et al., 2019) and do not exhibit uncrowding (Doerig, Bornet et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken et al., 2020), even when they are trained to ignore local features (Doerig, Bornet et al., 2020).

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

Downloaded from jov.arvojournals.org on 11/24/2021

multilevel perceptual phenomenon, as well as a precious and versatile probe to understand what may be missing from current models of human vision.

Keywords: crowding, grouping, face recognition, holistic processing, texture tiling model

Acknowledgments

The authors thank Ruth Rosenholtz for her detailed comments on this manuscript and for sharing the code of the TTM. We thank both reviewers for their insightful comments.

A.B. was supported by the European Union’s Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 785907 (Human Brain Project SGA2) and No. 945539 (Human Brain Project SGA3). O.H.C. was supported by the Swiss National Science Foundation (SNF) 320030_176153 “Basics of visual processing: from elements to figures.” A.D. was supported by the Swiss National Science Foundation grants No. 176153 “Basics of visual processing: from elements to figures” and No. 191718 “Towards machines that see like us: human eye movements for robust deep recurrent neural networks.” D.W. was supported by the National Institutes of Health grant R01 CA236793.

Author contributions: All the authors contributed to design the study. A.B., O.H.C., and M.M. conducted the experiments. A.B. and O.H.C. analyzed the data. A.B. and M.M. wrote the manuscript. O.H.C., A.D., D.W., and M.H.H. provided feedback on earlier drafts of the manuscript.

Consent for publication: All participants gave permission for the publication of their data.

Availability of data and material: All relevant data are available from the authors at alban.bornet@epfl.ch.

Commercial relationships: none.

Corresponding author: Alban Bornet.
Email: alban.bornet@epfl.ch.
Address: Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Switzerland.

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.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*(12), e1006613.

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

Banks, W. P., Larson, D. W., & Prinzmetal, W. (1979). Asymmetry of visual interference. *Perception & Psychophysics, 25*(6), 447–456.

Banks, W. P., & Prinzmetal, W. (1976). Configurational effects in visual information processing. *Perception & Psychophysics, 19*(4), 361–367.

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.

Bornet, A., Doerig, A., Herzog, M. H., Francis, G., & Van der Burg, E. (2021). Shrinking Bouma’s window : How to model crowding in dense displays. *PLoS Computational Biology, 17*(7), e1009187.

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.

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*, 585921.

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

Downloaded from jov.arvojournals.org on 11/24/2021

crowding. *Frontiers in Integrative Neuroscience, 8*, 73.

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.

Choung, O. H., Bornet, A., Doerig, A., & Herzog, M. H. (2021). Dissecting (un)crowding. *Journal of Vision, 21*(10):10, 1–20.

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), 628–638.

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.

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.

Egeth, H. E., & Santee, J. L. (1981). Conceptual and perceptual components of interletter inhibition. *Journal of Experimental Psychology: Human Perception and Performance, 7*(3), 506.

Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision, 16*(2), 13.

Faivre, N., & Kouider, S. (2011a). Increased sensory evidence reverses nonconscious priming during crowding. *Journal of Vision, 11*(13), 16.

Faivre, N., & Kouider, S. (2011b). Multi-feature objects elicit nonconscious priming despite crowding. *Journal of Vision, 11*(3), 2.

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.

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.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gong, M., Xuan, Y., Smart, L. J., & Olzak, L. A. (2018). The extraction of natural scene gist in visual crowding. *Scientific Reports, 8*(1), 1–13.

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.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques,* 229–238.

Herrera-Esposito, D., Coen-Cagli, R., & Gomez-Sena, L. (2021). Flexible contextual modulation of naturalistic texture perception in peripheral vision. *Journal of Vision, 21*(1), 1.

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.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object

Downloaded from jov.arvojournals.org on 11/24/2021

recognition : A matter of appearance. *Journal of Vision, 15*(6), 5.

Herzog, M. H., Sayim, B., Manassi, M., & Chicherov, V. (2016). What crowds in crowding? *Journal of Vision, 16*(11), 25.

Huckauf, A., Heller, D., & Nazir, T. A. (1999). Lateral masking : Limitations of the feature interaction account. *Perception & Psychophysics, 61*(1), 177–189.

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.

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., Bonneh, 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.

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.

Levi, D. M., Klein, S. A., & Hariharan, S. (2002). Suppressive and facilitatory spatial interactions in foveal vision : Foveal crowding is simple contrast masking. *Journal of Vision, 2*(2), 2.

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.

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

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.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *JOSA A, 7*(5), 923–932.

Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision, 15*(8), 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.

Manassi, M., Sayim, B., & Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision, 12*(10), 13.

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

Mason, M. (1982). Recognition time for letters and nonletters : Effects of serial position, array size, and processing order. *Journal of Experimental Psychology: Human Perception and Performance, 8*(5), 724.

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.

Mewhort, D. J. K., Marchetti, F. M., & Campbell, A. J. (1982). Blank characters in tachistoscopic recognition : Space has both a symbolic and a sensory role. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 36*(4), 559.

Downloaded from jov.arvojournals.org on 11/24/2021

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.

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.

Pittino, F., Eberhardt, L. V., Kurz, A., & Huckauf, A. (2019). Crowding with Negatively Conditioned Flankers and Targets. *Advances in Cognitive Psychology, 15*(1), 1.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–70.

Reuther, J., & Chakravarthi, R. (2019). Response selection modulates crowding : A cautionary tale for invoking top-down explanations. *Attention, Perception, & Psychophysics, 82*, 1763–1778.

Rosenholtz, R. (2014). Texture perception. *Oxford Handbook of Perceptual Organization, 167*, 186.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*, 437–457.

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.

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

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

Saarela, T. P., Westheimer, G., & Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *Journal of Vision, 10*(10), 17.

Sayim, B., Greenwood, J. A., & Cavanagh, P. (2014). Foveal target repetitions reduce crowding. *Journal of Vision, 14*(6), 4.

Sayim, B., Manassi, M., & Herzog, M. (2014). How color, regularity, and good Gestalt determine backward masking. *Journal of Vision, 14*(7), 8.

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Figural grouping affects contextual modulation in low level vision. *Journal of Vision, 8*(6), 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.

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.

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.

Sun, H.-M., & Balas, B. (2015). Face features and face configurations both contribute to visual crowding. *Attention, Perception, & Psychophysics, 77*(2), 508–519.

Tannazzo, T., Kurylo, D. D., & Bukhari, F. (2014). Perceptual grouping across eccentricity. *Vision Research, 103*, 101–108.

Downloaded from jov.arvojournals.org on 11/24/2021

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 Computer Biology, 6*(1), e1000646.

Vickery, T. J., Shim, W. M., Chakravarthi, R., Jiang, Y. V., & Luedeman, R. (2009). Supercrowding : Weakly masking a target expands the range of crowding. *Journal of Vision, 9*(2), 12.

Wallace, J. M., Chiu, M. K., Nandy, A. S., & Tjan, B. S. (2013). Crowding during restricted and free viewing. *Vision Research, 84*, 50–59.

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.

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

Waugh, S. J., & Formankiewicz, M. A. (2020). Grouping Effects on Foveal Spatial Interactions in Children. *Investigative Ophthalmology & Visual Science, 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.

Xia, Y., Manassi, M., Nakayama, K., Zipser, K., & Whitney, D. (2020). Visual crowding in driving. *Journal of Vision, 20*(6), 1.

Yeshurun, Y., & Rashal, E. (2010). Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision, 10*(10), 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.

Ziemba, C. M., & Simoncelli, E. P. (2021). Opposing effects of selectivity and invariance in peripheral vision. *Nature Communications, 12*(1), 4597.

Downloaded from jov.arvojournals.org on 11/24/2021

Supplementary information

Suppl. Inf. A:

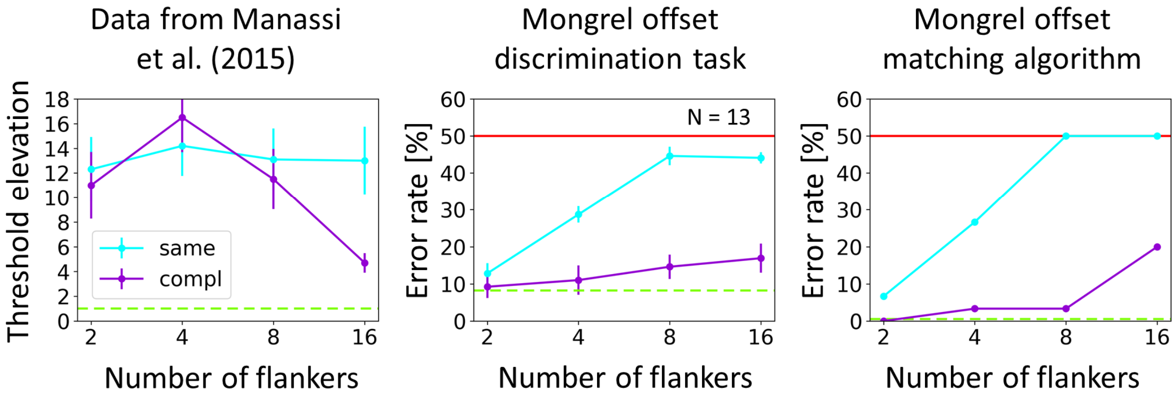


Figure A. Comparison between Lines and Completion experiments. Left. Offset discrimination thresholds were determined for vernier targets presented in peripheral vision. Bars indicate flanker configurations threshold elevation compared to the vernier alone (green dashed line). Data are taken from Manassi 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.

Suppl. Inf. B:

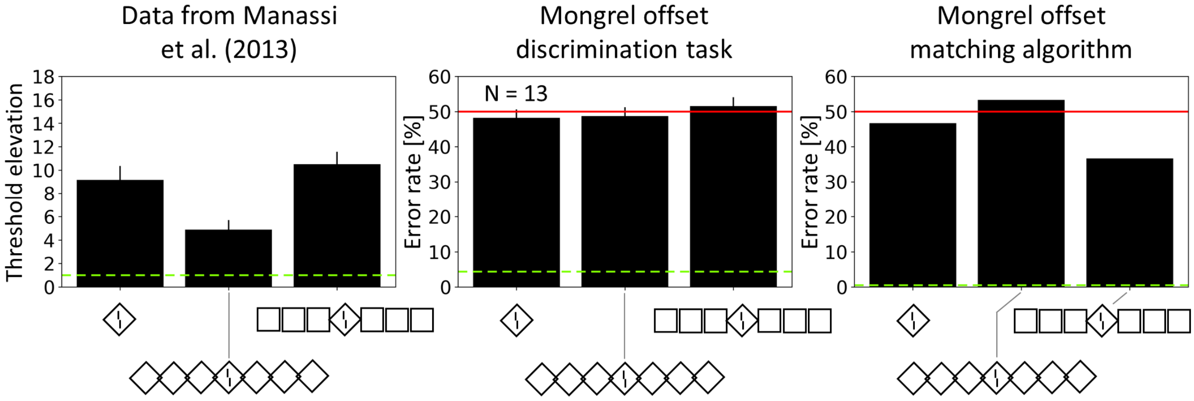


Figure B. Shapes experiment with diamonds. Left. Offset discrimination thresholds were determined for vernier targets presented in peripheral vision. Bars indicate flanker configurations threshold elevation compared to the vernier alone (green dashed line). Data are taken from Manassi et al. (2013). **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 (Fig. C, Suppl. Inf.).

Suppl. Inf. C:

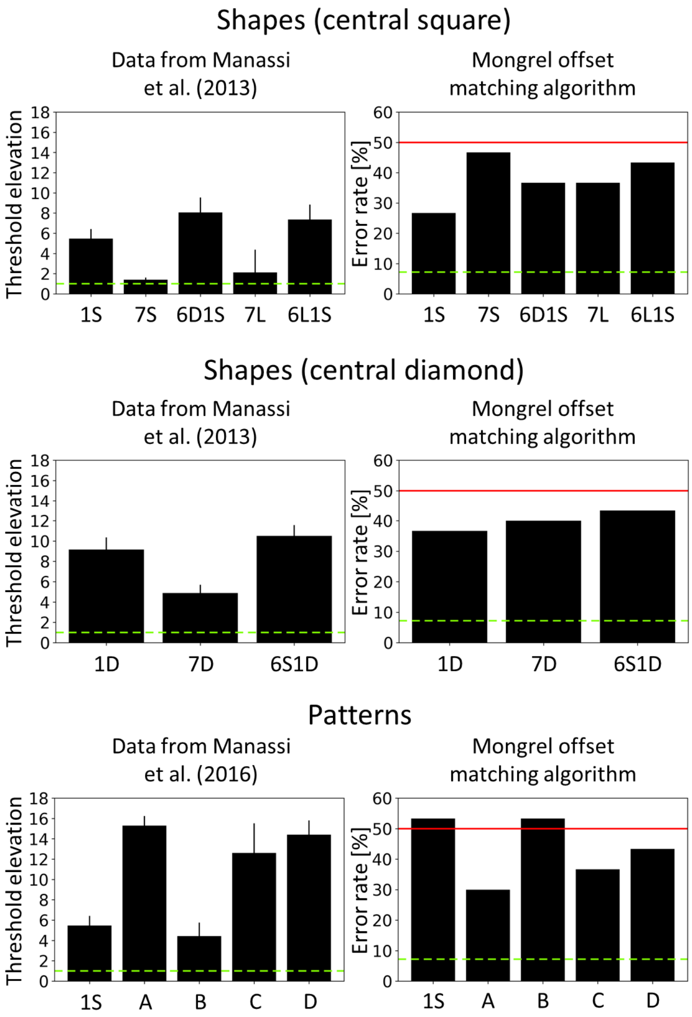


Figure C. Shapes and Patterns experiments with a larger fovea radius parameter in the TTM. Left. Offset discrimination thresholds were determined for vernier targets presented in peripheral vision. Bars indicate flanker configurations threshold elevation compared to the vernier alone (green dashed line). 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 Figure B Suppl. Inf..

Suppl. Inf. D:

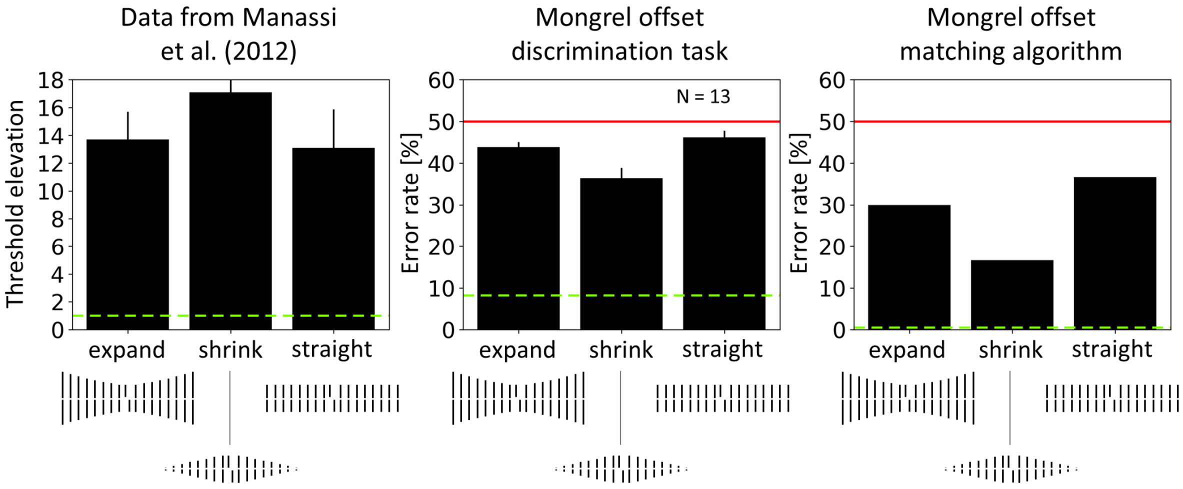


Figure D. Butterflies. Left. Data from (Manassi et al., 2015). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. **Center.** TTM validation in which observers discriminate between left and right offset verniers in mongrel images. **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). Note that the y-axis labels are different.

Suppl. Inf. E:

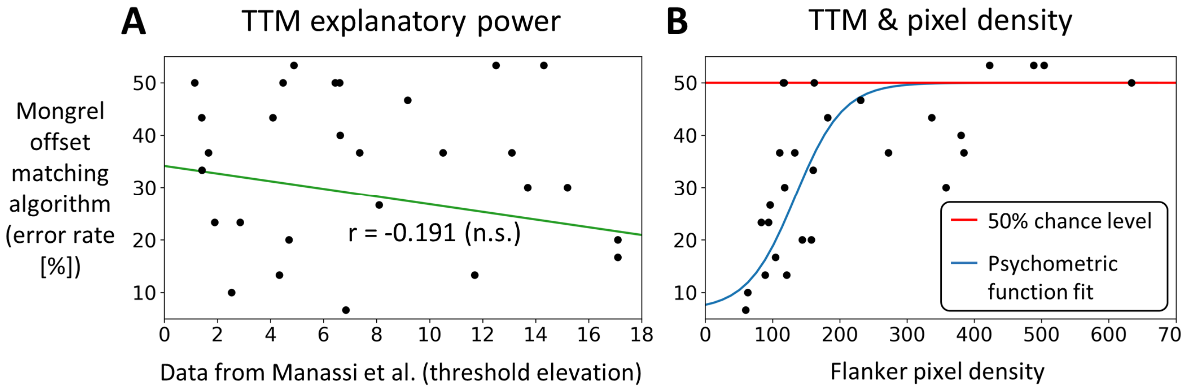


Figure E. A. When plotting error rates in the mongrel offset matching algorithm as a function of psychophysical thresholds data from Manassi et al. (2012, 2013, 2015, 2016), no correlation was found (r(36)=-0.191, p=-0.264, $BF_{01}=2.647$). **B.** We plotted the error rates measured in the mongrel offset 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, $BF_{10}>10^4$).

Suppl. Inf. F:

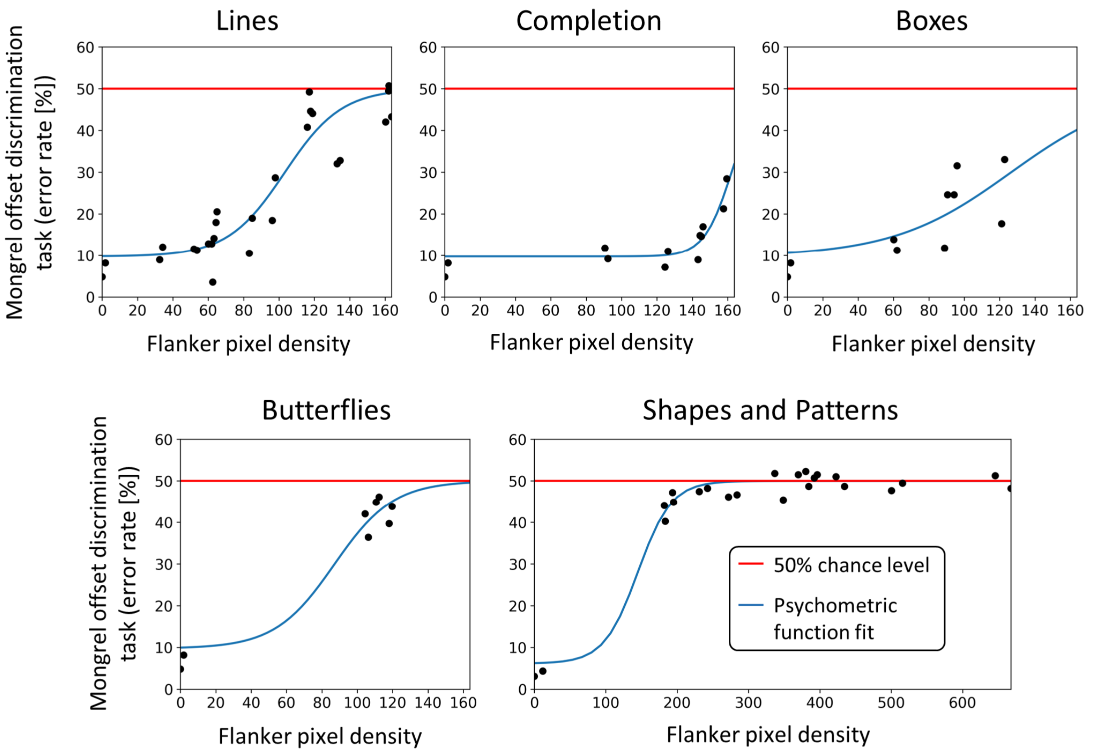


Figure F. We plotted the error rates measured in the mongrel offset discrimination task with all tested flanking conditions as a function of the sum of the flanker pixel density (see Methods for details). Each dot indicates a flanking condition in Figure 1. The red line indicates chance level performance. 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.

Suppl. Inf. G:

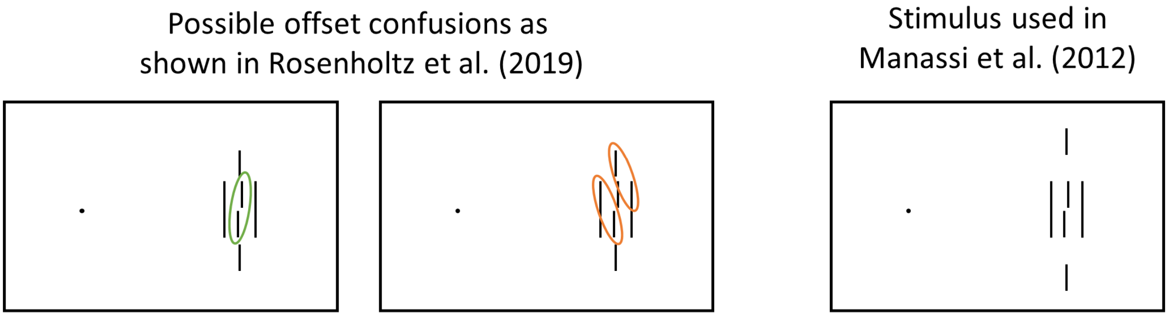

Possible offset confusions as shown in Rosenholtz et al. (2019)
Stimulus used in Manassi et al. (2012)

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 instead increase crowding by creating multiple offsets among vernier, flankers and pointers lines, because of the location uncertainty of each element in the visual field (Rosenholtz et al., 2019). **Left.** Possible representation of the stimulus used in the crowding experiment of Manassi et al. (2012), as depicted in Rosenholtz et al. (2019). Participants may report the correct vernier offset (green) or mistakenly report another offset, because of location uncertainty (orange). In this case the pointers are reconstructed very close to the vernier target. **Right.** The pointers used in the actual experiment are quite far from the vernier target, making the location uncertainty argument unlikely, or responsible for very few substitution errors. You can convince yourself simply by looking at the fixation point and checking whether you would easily confuse one of the pointers for a fragment of the target.

Suppl. Inf. H:

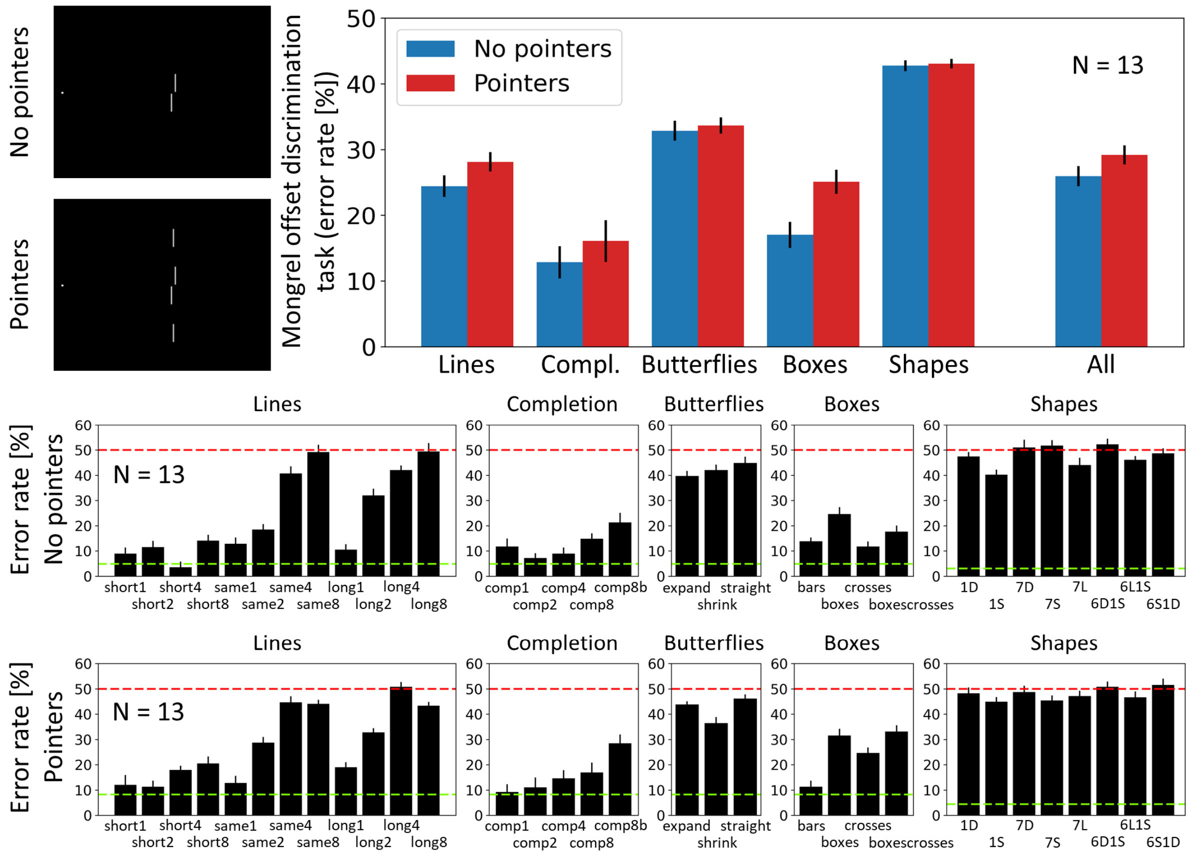


Figure H. We measured human performance in the mongrel offset discrimination task for all conditions in Manassi et al. (2012, 2013, 2015), with or without pointers (bottom). The actual layout of the different conditions is shown in Figure 1. The TTM did not show any significant increase in crowding strength (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-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).

Suppl. Inf. I:

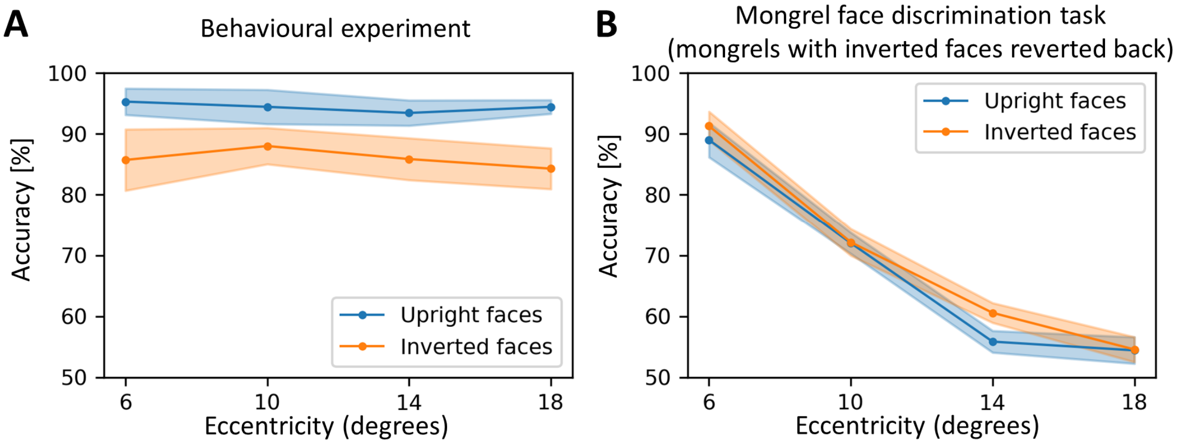


Figure I. TTM & single Mooney face discrimination, reverted-back version. A. Single face discrimination task. Copied from Figure 11A for comparison. Observers were able to discriminate an upright or an inverted face from a scrambled face at all tested eccentricities. Moreover, performance was higher 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 11B, 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 (χ^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: $r_m^2=0.815$, $r_c^2=0.822$).

Suppl. Inf. J:

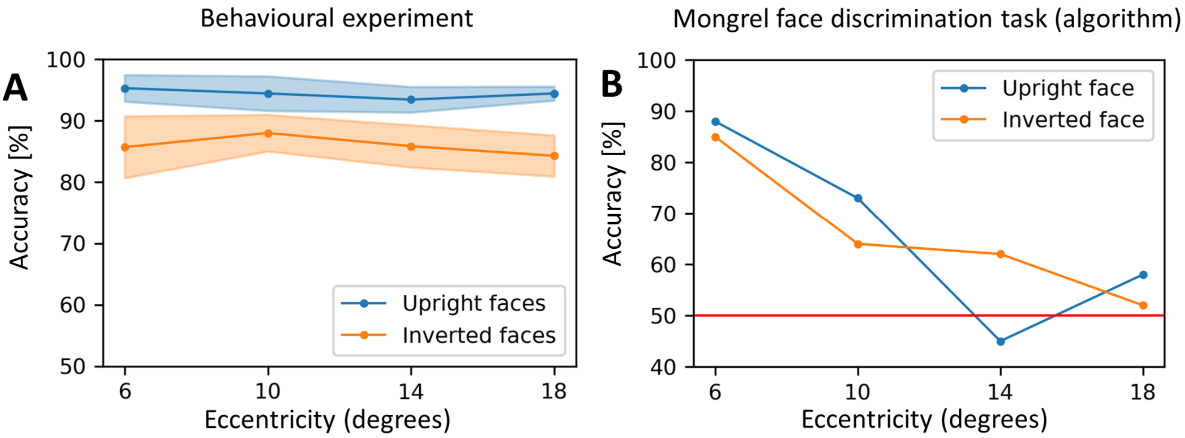


Figure J. TTM & single Mooney face recognition, algorithm results. A. Face discrimination task. Copied from Figure 11A for comparison. Observers were asked to discriminate an upright/inverted face from a scrambled face at all tested eccentricities. Accuracy remained on a constant high level for all eccentricities. Crucially, accuracy was higher for upright than for inverted faces. **B.** Mongrel face discrimination task, algorithm results. The results are qualitatively similar to the mongrel face discrimination task (Figure 11B). Accuracy decreased with increasing eccentricity and no notable difference between the upright and inverted face conditions was observed, contrary to the behavioral results.

Suppl. Inf. K:

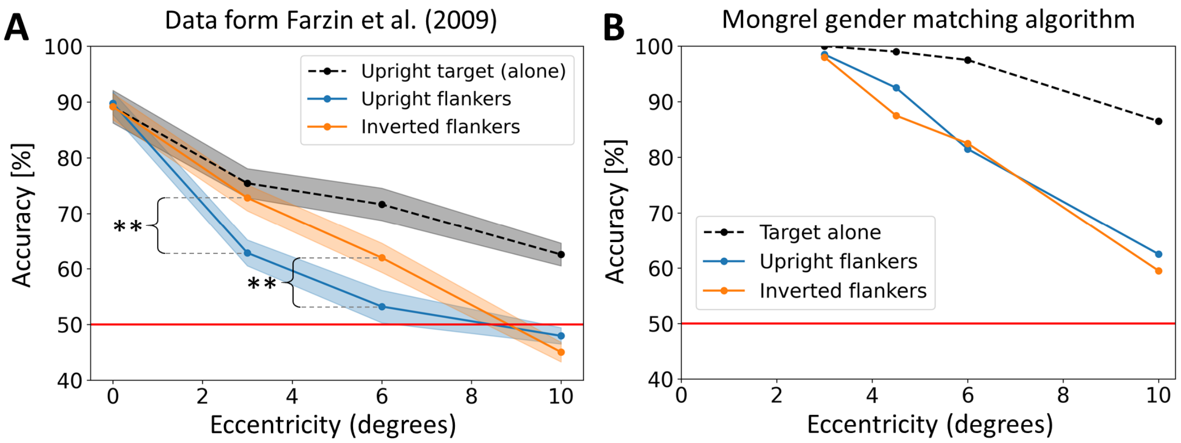


Figure K. TTM & crowding in Mooney faces, algorithm results. A. Face crowding task, data from Farzin et al. (2009). Copied from Figure 12A for comparison. Target discrimination performance decreased when eccentricity increased. When the target face was flanked by inverted faces, crowding increased with 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.

Suppl. Inf. L:

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%).

Appendix B

Machiraju, H.*, **Choung, O. H.***, Frossard, P., & Herzog, M. H. (2021). Bio-inspired Robustness: A Review. *arXiv preprint arXiv:2103.09265*.

Bio-inspired Robustness: A Review

Machiraju, Harshitha[1,2]* Choung, Oh-Hyeon[1]* Frossard, Pascal[2] , Herzog, Michael. H.[1]

[1]Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, http://lpsy.epfl.ch, harshitha.machiraju@epfl.ch, oh-hyeon.choung@epfl.ch

[2]Signal Processing Laboratory 4 (LTS4), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, https://www.epfl.ch/labs/lts4/

* The authors contributed equally.

Abstract

Deep convolutional neural networks (DCNNs) have revolutionized computer vision and are often advocated as good models of the human visual system. However, there are currently many shortcomings of DCNNs, which preclude them as a model of human vision. For example, in the case of adversarial attacks, where adding small amounts of noise to an image, including an object, can lead to strong misclassification of that object. But for humans, the noise is often invisible. If vulnerability to adversarial noise cannot be fixed, DCNNs cannot be taken as serious models of human vision. Many studies have tried to add features of the human visual system to DCNNs to make them robust against adversarial attacks. However, it is not fully clear whether human vision inspired components increase robustness because performance evaluations of these novel components in DCNNs are often inconclusive. We propose a set of criteria for proper evaluation and analyze different models according to these criteria. We finally sketch future efforts to make DCCNs one step closer to the model of human vision.

1. Introduction

Deep convolutional neural networks (DCNN) have revolutionized computer vision. DCNNs reach near or even super-human performance in many tasks, such as image classification (He et al., 2016), image segmentation (He et al., 2017), image captioning (Karpathy & Fei-Fei, 2015), and image generation (Choi et al., 2020). There is now a fierce debate whether DCNNs are also a good model for the human visual system. On the one hand, proponents argue that DCNNs perform like humans in many object recognition tasks, and their architecture indeed resembles the human one (Kubilius et al., 2018; Schrimpf et al., 2020). On the other hand, DCNNs often solve vision tasks very differently than humans (Geirhos et al., 2020), indicating that comparable performance levels *per se* do not tell whether DCNNs are good models.

In this contribution, we look at this pro-con dichotomy from the perspective of robustness, that is, the ability of DCNNs to make proper object classifications even when data are slightly perturbed or coupled with noise. Specifically, we consider adversarial attacks, which present a major problem for DCNNs (Sharif et al., 2016; Y. Zhang et al., 2019). These attacks are small, crafted perturbations that cause major misclassifications of the DCNNs even though they are imperceptible to humans (Szegedy et al., 2014). Figure 1 shows two adversarial examples where the image of a dog with imperceptible noise is misclassified as red wine or toilet paper by DCNNs. Obviously, humans and DCNNs show very different behavior, hence at first glance, DCNNs should be discarded as proper models for human vision. However, it may be that minor fixes can make DCNNs robust to adversarial attacks, for example, by taking inspiration from human vision models. Such attempts may help to bridge the performance gap that still exists between DCNNs and human vision.

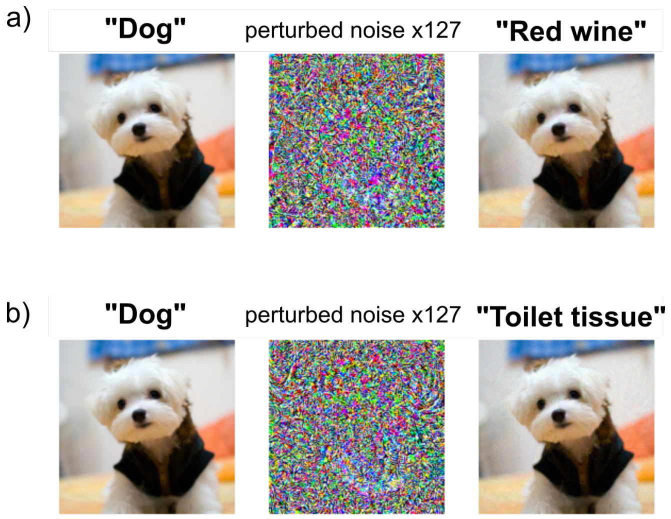


Figure 1. Adversarial examples, using PGD with L_∞ *and with noise constraint of* $\epsilon = 16/255$ *on 'n02085936_6883.jpeg' of ImageNet (Deng et al., 2009) dataset. The left image is the original, and the right image is the original plus the noise image shown in the middle. For humans, the differences between the original image and the original plus noise image are hardly visible. For DCNNs, the noise leads to serious misclassification.*

Many investigations have proposed improvements against adversarial attacks from both the Computer vision (e.g., Athalye et al., 2018b; Carlini et al., 2019; Croce et al., 2020; Tramer et al., 2020) and the Neuroscience communities (e.g., Choksi et al., 2020; Kiritani & Ono, 2020; Marchisio et al., 2020; Rusak et al., 2020; Zoran et al., 2020). Improving robustness is crucial not only to reduce vulnerability to adversarial attacks but also for improving the transfer of learning (Salman et al., 2020; Utrera et al., 2020), image segmentation (Salman et al., 2020), generalization (Bochkovskiy et al., 2020; Song et al., 2020; Xie et al., 2020), etc. (see Fig. 2). In this paper, we focus on bio-inspired methods, with the main objective to establish stronger connections between DCNNs and human vision.

Unfortunately enough, different bio-inspired approaches trying to protect DCNNs against adversarial attacks by including components of the biological visual system have often reached unresolved conclusions. We advocate that one of the main reasons for this situation is the non-uniform ways of evaluating and analyzing these components.

In this paper, we try to remedy this situation by proposing criteria to standardize the evaluation and analysis methods that each study needs to meet. We first review definitions of robustness and adversarial attacks, and evaluation methods in Section 2. We then explore different studies and summarize their results and analysis in Sec 3 and 4. Finally, we summarize the main learning messages and propose new insights for future research directions towards understanding better how the joint development of DCNN architectures and human vision models could lead to cross-fertilization, which could for example lead to better and more robust computer vision systems.

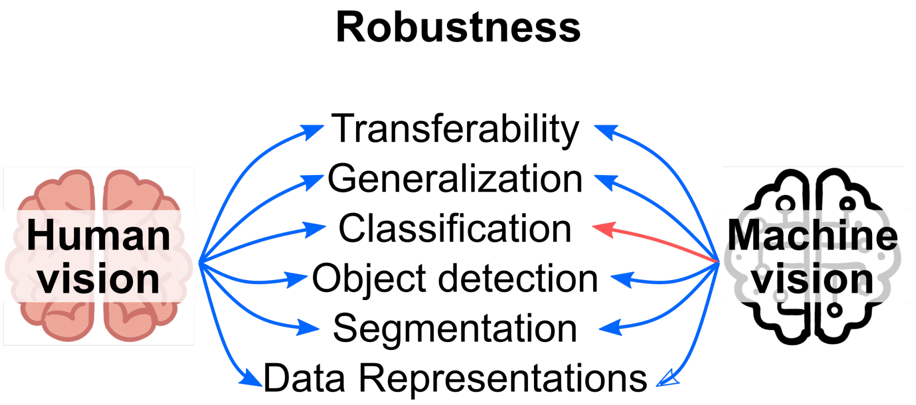


Figure 2. Robust features are transferable to other tasks. Since most Deep Learning models are solely optimized to be highly accurate for a given task, they do not generalize to other tasks, unlike humans (Lapuschkin et al., 2019). Ilyas et al., 2019 showed that by design, Deep Learning models tend to use highly predictive, non-robust features (which include background, texture/high-frequency information, etc.) instead of robust features (which include foreground, shape/low-frequency information; Zhang & Zhu, 2019, etc.). This explains, to some extent, the high vulnerability of DCNNs to adversarial attacks. Hence, by forcing DCNNs to utilize robust features, we may obtain more human-vision-like representations (Engstrom et al., 2019; Kaur et al., 2019; Tsipras et al., 2018). Since robust networks utilize global information like shapes, they have a better "understanding" of the overall features of each class, which improves generalization to unseen distributions (Bochkovskiy et al., 2020; C. Song et al., 2020; Cihang Xie et al., 2020), better image segmentation and object detection abilities (Salman et al., 2020), and adaptation across domains since the features learned are now generic for each class (transfer learning; Salman et al., 2020; Utrera et al., 2020).

2. Adversarial Robustness

2.1. Adversarial Attacks and Defenses

In adversarial attacks, carefully chosen noise added to images containing objects leads to gross misclassification of these objects (Szegedy et al., 2014, Fig. 1). Interestingly, the noise is usually imperceptible for humans. For example in Fig .1, humans can hardly distinguish the original image with the dog and the same image with additive noise. Adversarial attacks are mathematically defined as follows: for a given image x, and a classifier $f: \mathbb{R}^n \rightarrow \mathbb{R}^c$, where $x \in \mathbb{R}^n$ and y is the **true** class label we define the adversarial noise Δx as:

$$\mathit{arg\ max}\ (f(x + \Delta x)) \neq \mathit{arg\ max}\ (f(x))\ \text{s.t}\ ||\ \Delta x\ ||_p \leq \epsilon \qquad \text{(Eq. 1)}$$

The true label class (y) that the input (x) belongs to is referred to as the *source class,* and the *target class* is the class to which the adversarial sample ($x + \Delta x$) is misclassified by the DCNN model. In the above equation, the upper bound on the noise ϵ is added to ensure that the noise remains small. In Eq. 1, Adversarial perturbations are defined w.r.t. an L_p norm. The usual values for *p* are *0, 1, 2, and* ∞ can be summarized as follows:

- L_0 norm: Constrains the number of non-zero pixels in the adversarial perturbation (sparsity).
- L_1norm (Manhattan distance): Constrains the absolute value of the magnitude of the adversarial perturbation.
- L_2 norm (Euclidean distance): Constrains the squared value of the magnitude of the adversarial perturbation.
- L_∞ norm: Constrains the maximum element of the adversarial perturbation.

The upper bound for adversarial perturbations, ϵ is chosen w.r.t. an L_p norm (Fezza et al., 2019; Jordan et al., 2019) such that the noise is imperceptible to humans (Bouniot et al., 2021).

Adversarial perturbations in Eq. 1 may be defined w.r.t. any norm or distance metrics other than the L_p norm metrics (Jordan et al., 2019). But, the L_p metrics are most studied, well understood, and are much simpler to use than other distance metrics (Akhtar & Mian, 2018; Ortiz-Jimenez et al., 2020). Hence throughout this paper, we focus on $L_p$norm based adversarial attacks and robustness.

There are different algorithms to generate adversarial attacks, including the Fast Sign Gradient Method (FGSM; Szegdy et al., 2014), Projected gradient descent (PGD; Madry et al., 2017), Jacobian Based Saliency Map Attacks (JSMA; Papernot et al., 2016), DeepFool (Moosavi-Dezfooli et al., 2016), Decision boundary attacks (DBA; Brendel et al., 2018), Carlini-Wagner attacks (Carlini & Wagner, 2017), Basic iterative method (BIM; Kurakin et al., 2017), etc. More information can be found in reviews such as Akhtar & Mian, 2018 and Ortiz-Jimenez et al., 2020.

DCNNs, unlike humans, do not have any understanding of object features. DCNNs learn discriminative features that are maximally capable of providing high accuracy on the trained data and construct decision boundaries based on these features. These learned features focus on local information of the image, such as texture (Song et al., 2020). It is this over-reliance on local features that makes DCNNs an easy prey for small adversarial noises (Ilyas et al., 2019).

The machine learning community has explored different ways to make DCNNs more robust (defend) against adversarial perturbations. Adversarial training (Madry et al., 2017; Moosavi-Dezfooli et al., 2016) is regarded as the state-of-the-art defense method. This method trains DCNNs by utilizing adversarial samples generated by attacks and augmenting the training dataset with these samples. By doing so, the DCNN is forced to change its decision boundaries to make sure that the adversarial samples belong to the same class as the original input sample. However, this type of training comes with a large computational load. For each input sample, the DCNN must first create adversarial perturbations via attacks and then learn to classify them correctly. Considering humans do not have to follow a similar procedure, there might be a simpler way to make DCNNs robust to adversarial perturbations. Many works have tried to use human vision-like features for creating more robust DCNNs (Chakraborty et al., 2018). These studies mainly focus on changing the decision boundaries of the DCNNs by learning global rather than local features similar to human vision (Engstrom et al., 2019).

However, as shown by Carlini et al., 2019 due to the failure to systematically evaluate the robustness of a defense, it is unclear how good these methods are (Brendel & Bethge, 2017 vs. Nayebi & Ganguli, 2017).

2.2. Evaluation of Robustness

2.2.1 Evaluation

There are different ways to measure the robustness of a DCNN against adversarial attacks, including:

Adversarial Accuracy: Defined as the fraction of correctly classified adversarial samples out of all adversarial samples created on a dataset:

$$\text{Adversarial Accuracy} = \frac{\#correctly\ classified\ adversarial\ samples}{\#adversarial\ samples} \quad \text{(Eq. 2)}$$

Most work uses this metric to measure robustness.

Attack success rate: Defined as the fraction of adversarial samples generated by a given attack that succeeds in fooling the DCNN model out of all the generated adversarial samples.

$$\text{Attack success rate} = \frac{\#misclassified\ adversarial\ samples}{\#adversarial\ samples} \quad \text{(Eq. 3)}$$

Mean Distortion: Defined as the median distance of all pairs built on a source image in a given dataset (D) and the corresponding adversarial sample.

$$\text{Mean distortion} = median\ (\sum_{\forall x \in D} ||\ x_{adv} - x\ ||_p)\ \text{where } p \in \{0, 1, 2, \infty\} \quad \text{(Eq. 4)}$$

In general, the evaluation of robustness heavily depends on the attack used for generating the adversarial samples. The adversarial attack itself depends on parameters such as the number of iterations, the ϵ values, step sizes, etc. Typically, as many adversarial attacks are constructed on the estimation of gradients in Eq. 1, one has to make sure that the evaluation really captures the intrinsic robustness of the model, and not merely the impossibility to compute adversarial examples through specific, gradient-based methods, referred to as *gradient masking*.

DCNNs learn through backpropagation and minimize the classification loss by using the gradient of the loss function. Adversarial perturbations are created by maximizing the classification loss to cause misclassification. Thus, adversarial perturbations need gradients to maximize the classification loss.

Gradient masking refers to hiding this gradient information by using non-differentiable operations like subsampling (Hosseini et al., 2019), input transformations (Guo et al., 2018), or stochasticity (Dhillon et al., 2018). Gradient masking does not lead to adversarial robustness, it only makes it difficult for gradient-based adversarial attacks to find adversaries (Athalye et al., 2018). If an attack succeeds in finding a hidden adversarial sample (despite gradient masking), the DCNN still ends up misclassifying it. Gradient masking does not improve the decision boundary, it just avoids existing gradient-based adversarial attacks.

Given the high variability of methods for the evaluation of adversarial robustness, for any defense algorithm that aims at improving robustness, it is necessary to follow the basic defense evaluation guidelines provided by Carlini et al., 2019. Specifically, we stress the importance of checking for gradient masking phenomena, as explained above.

2.2.2 Bio-inspired robustness evaluation

As explained earlier, if bio-inspired defenses use gradient masking, they are not truly making DCNNs functioning closer to human visual perception. Hence, bio-inspired robustness methods need to ensure that they are not using gradient masking. Here, we summarize the sanity checks to identify gradient masking (for 1-4 Athalye et al., 2018):

1. **One-step attacks perform better than iterative attacks:** When an adversarial attack comes with multiple iterations (Madry et al., 2017), it has a better chance to succeed than an attack with a single iteration (Szegedy et al., 2014), e.g., PGD (multi-iteration attack) vs. FGSM (1 iteration attack). Failure to see this usually indicates gradient masking.
2. **Observing irregularities in perturbation budget curves:** Perturbation budget refers to the maximum value of ϵ used to generate adversarial examples as in Eq. 1. The higher the perturbation budget, the better the chance of attack success. Plots of the Attack success rate vs. perturbation budget are called perturbation budget curves. Ideally, the attack success rate should increase with the perturbation budget (larger noise) and for extremely high values of ϵ the attack success rate would be 100%. If this behavior is not seen in the perturbation curve, there is likely gradient masking.
3. **Random sampling finds adversarial perturbations:** Adversarial attacks are optimized to find adversarial perturbations and hence have a much higher chance of causing misclassification than any randomly sampled perturbation. If the success rate of adversarial attack is lower than random perturbations there is likely gradient masking.
4. **Black-box attacks are better than white-box attacks:** In order to generate adversarial perturbations, most attacks typically need information about the model being used, especially its parameters. When attacks are generated with complete access to model information, they are called white-box attacks. Attacks without any access to the model and its parameters are called black-box attacks. Black-box attacks usually have a lower attack success rate than white-box attacks. Failure to see this usually indicates gradient masking.

5. **Non-gradient-based adversarial attacks:** Other than the above suggested by Athalye et al., 2018, non-gradient-based adversarial attacks like Decision Boundary attacks (DBA; Brendel et al., 2018) can also be used to identify gradient masking. Let us consider the robustness of a standard model (without any defense method) evaluated on a non-gradient-based attack to be r_s, and for the defended model the robustness is r_d. If the defense uses gradient masking, it will be broken by the non-gradient-based attack, since masked (hidden) gradient information makes no difference to such attacks. Hence, for defenses with gradient masking, $r_d \approx r_s$. If there is no gradient masking $r_d > r_s$.

Since the use of non-differentiable layers or stochasticity could result in gradient masking, the above checklist can be used to validate the evaluation of the actual robustness of a model. Additionally, the adversarial attacks proposed by Athalye et al., 2018, which are specially designed to break gradient masking based defenses, can also be used to evaluate defenses with similar components.

3. Biologically Inspired Components

We review a series of methods, which propose to increase robustness against adversarial attacks by adding features of the human visual system to DCNNs. We first review some design guidelines, and then we highlight specific bio-inspired components and their properties. These methods were selected from the recent machine learning literature and are summarized in Table 1.

3.1. Choice of Bio-Inspired Components

In bio-inspired robustness, the idea is to make DCNN robust like humans while employing the concepts used by the latter. One problem is that it is rather unclear how the visual system acquires robustness. Hence, when implementing a new component, it is important to motivate, analyze, and validate the suggested components. Accordingly, when applying bio-inspired components for robustness, evidence should be provided that these components actually improve the DCNN.

In particular, it is necessary to test whether the contribution of each component indeed increases robustness. We can do so by “freezing” the component and analyze the change in robustness in comparison when the component is active. Alternatively, we can use simpler variants of the bio-inspired component.

Finally, it is also important to validate the design choices by verifying that the proposed bio-inspired component is indeed increasing robustness for the reasons it was motivated. For example, if Gaussian filtering is thought to improve robustness because it filters out adversarial noises, the authors should show that indeed adversarial noise is filtered out, on top of other robustness measures (Xie et al., 2019). Methods for validation include weight visualization, representation visualization, saliency maps, etc.

We review below different recent studies, which have proposed to include bio-inspired components in order to improve the robustness of DCNNs, and discuss motivations, design choices, and expected benefits.

3.2. Early visual processes

In human vision, the visual inputs are processed in the following order:

Retina → Lateral Geniculate thalamic Nucleus (LGN) → Primary Visual Cortex (V1)

The retina has a higher density of photoreceptors in the fovea than in the periphery. This uneven distribution of photoreceptors results in a non-linear sampling of the visual input. Therefore, the image resolution is increased in the fovea and blurred towards the periphery. This, in turn, increases the Signal-to-Noise Ratio (SNR) of the input, which was proposed to increase robustness (Elsayed et al., 2018).

Retinal neurons use lateral inhibition, which attenuates redundant and irrelevant signals (decorrelation; Segal et al., 2015). Thus, only significant signals (Bakshi & Ghosh, 2017), which are required to properly classify the object, survive. Thus, we suggest lateral inhibition may be useful for increasing robustness.

The Lateral geniculate nucleus (LGN), then modulates output signals from the retina using feedback signals from higher visual areas (Usrey & Alitto, 2015), to ensure better classification.

In the primary visual cortex (V1), neural processing is similar to a bank of Gabor filters (GFB) with multiple orientations, spatial frequencies, etc. (De Valois, Albrecht, et al., 1982; De Valois, William Yund, et al., 1982; Ringach, 2002). Due to its diversity, the GFB decomposes the signal into a large number of disentangled features. These features pass through either simple or complex cells. Simple cells have a linear response and discard irrelevant information. Complex cells have a non-linear response to detect more complex features (Vintch et al., 2015). It may be that features generated by GFB, simple cells, and complex cells are necessary for downstream areas to increase robustness (Dapello et al., 2020).

In V1, multiple layers of neurons are organized as cortical minicolumns. Each minicolumn receives inputs from the same receptive field and all minicolumns have the same receptive field *size*. Thus, each minicolumn is similar to a vector encoding features like pose, orientation, scale, etc. (Hinton, 1981; Hubel & Wiesel, 1963). This preserves the spatial relationships within an object, which is crucial for downstream object perception , and, thus, we suggest that it may increase robustness.

From V1 to V4, pooling takes place (Freeman & Simoncelli, 2011) causing an increase in receptive field sizes and hence creating different scales across the visual stream. Pooling may help to reduce dimensions by removing irrelevant information and creating abstractions of the object (Poggio et al., 2014), and, thus, we suggest that it may increase robustness.

Functionally, the entire cortex is known to encode the input signals in a sparse way, which improves the selectivity of the class-relevant features (Paiton et al., 2020), and, thus, we suggest that it may increase robustness.

Additionally, all neuronal processes always include stochasticity, and such stochasticity is thought to contribute to the generalization of the signals (Echeveste & Lengyel, 2018), and, thus, we suggest that it may increase robustness.

Among the works that take inspiration from early vision models to improve the robustness of DCNNs, we can first outline the study of **Reddy et al., 2020**, which implemented two **sampling** methods from biological vision. One is the non-uniform sampling of the retina, and the other is the multi-scale sampling (due to pooling) of the cortex. The authors implemented both methods similarly, as briefly illustrated in Fig. 3. S1. First, samples were sampled with either method. Then each sample was processed by a DCNN. The outputs of all the processed samples were averaged to obtain the final classification output. The authors trained the DCNN with these sampling methods to increase robustness.

Then, **Dapello et al., 2020** imitated **neuronal features of V1**. The authors prefixed the DCNN with their custom model of V1, called VOneBlock. It consists of the GFB with parameters picked from empirical distributions (De Valois, Albrecht, et al., 1982; De Valois, William Yund, et al., 1982; Ringach, 2002); simple cell linearity implemented with ReLU and complex cell non-linearity implemented with quadrature phase-pair spectral power (Carandini et al., 1997); neuronal stochasticity with Poisson distribution parameters obtained from primate V1 (Softky & Koch, 1993). The authors trained the DCNN prefixed with their VOneBlock to increase robustness.

3.3. Feedback

Throughout the cortical visual system, information is processed through feedforward and feedback connections (Pennartz et al., 2019). Through feedback connections, higher layer contextual information modulates the activation patterns in the lower layers. Thus, it is implementing strong long-range spatial dependencies, global information extraction, perceptual grouping (Kreiman & Serre, 2020), and recognition of challenging images (Kar et al., 2019; Kietzmann et al., 2019). Since feedback encourages the use of more global information, we suggest that it may increase robustness (Elsayed et al., 2018; Olshausen, 2013).

One possible way of implementing feedback is Predictive Coding (Aitchison & Lengyel, 2017), which proposes that the brain continuously updates itself based on the prediction error between the input signal and its prediction.

We note the following studies that build on bio-inspired feedback mechanisms to improve robustness. **Huang et al., 2020** used predictive coding to construct a DCNN with feedback. Feedback is implemented with a modified Deconvolutional Generative Model (DGM; Nguyen et al., 2019). DGM takes the output from layer $(l+1)$of the DCNN, and then deconvolutes (transpose of convolution; reverse of convolution) to generate (predict) the output of layer(l). Input images h, are processed by the feedforward DCNN to produce intermediate representations $z$and predict the output label y. In the feedback pass, the predicted labels $y$are used to reconstruct the intermediate representations, $z,$ and then eventually to reconstruct the input as $\hat{h}$. As illustrated in Fig. 3. S4, the reconstruction ($\hat{h}$) of the input image (h), intermediate latent representation (z), and predicted output label (y) are dynamically modulating each other through feedforward and feedback processes. The authors trained their DCNN+feedback model using classification and reconstruction based losses to increase robustness of the DCNN.

Then, Capsule Networks (CapsNets) by Sabour et al., 2017 implement cortical minicolumn-like structure and also a predictive coding based feedback process. A Capsule is a vector of neurons that captures object features as well as its instantiation parameters, such as pose, orientation, lighting, etc (Fig 3. S3). CapsNets implement feedback using Routing by Agreement. The weights between the capsules in layer $(l+1)$and capsules in layer$(l)$are updated based on how well the lower layer capsule is able to predict the output of the higher layer capsule. This is done by finding the correlation (agreement) between the predictions of the higher and lower layer capsules. If the correlation is very high then the weight is increased (excitatory connection) else the weight is decreased to suppress irrelevant capsules (inhibitory connection). CapsNets use classification and reconstruction losses to update themselves.

Qin et al., 2019 found that since CapsNets bear a very high resemblance to human representations (Sabour et al., 2017), it takes a very large amount of adversarial noise to misclassify them. When large adversarial noise is added to the image, it resembles the target class more than the original/ source class of the image (Santurkar et al., 2019). This causes the reconstructed image to resemble the target class more than the original class of the input. The authors used this discrepancy between the original and the reconstructed image for the successful detection of adversarial perturbations. For a given input sample x, the CapsNet reconstructs it using the predicted class information, as x'. Then the L_2(euclidean) distance between them is found as $d(x, x')$. If this distance exceeds a preset threshold, then the input x is said to be adversarial. Else, it is declared a normal image. The authors showed that their CapsNet based detection is successful in the detection of adversarial samples.

Finally, **Kim et al., 2020** implemented the entire visual system (i.e., retina, LGN, V1-V4, feedback and lateral inhibition, neuronal stochasticity, and sparse coding) as closely as possible. The authors then prefixed it to a DCNN to increase robustness.

3.4. Miscellaneous

In addition to early vision and feedback, other bio-inspired components have been studied recently. Actually, it has been argued that neurophysiological data itself may contain robust representations that are used in the human visual system. Thus, as a proof of concept, **Li et al., 2019** regularized DCNNs using neurophysiological data obtained from mice V1. The authors first measured the mouse V1 neuronal responses for each image and then computed the response similarity matrix for the image pairs. Then, the similarity matrix was used to regularize the DCNN to increase robustness.

In addition, sleep is thought to be significant for the retention of memory. The short-term memory content is transmitted to long-term memory when sleeping (Rasch & Born, 2013), i.e., memory consolidation. Memory consolidation (Rasch & Born, 2013) reduces overfitting and improves generalization (Lewis & Durrant, 2011; Wamsley et al., 2010; González et al., 2020; Wei et al., 2018), thus, we suggest it may increase robustness. **Tadros et al., 2020** thus implemented a sleep-like algorithm by using Spiking Neural Networks (SNNs; Diehl et al., 2015). The authors used SNNs and spike time-dependent plasticity (STDP; Song et al., 2000) to implement the sleep-like algorithm. They first trained the DCNN and then transformed it into an SNN. Then the training images were fed into the SNN to induce the reactivation of the neurons, which leads to STDP. STDP enables the weights of highly related neurons to be strengthened and weakly related neurons to be weakened, to increase robustness.

Table 1. Summary of studies. We refer to each paper as S# instead of cross-referencing. 'Components' refers to the bio-inspired feature the authors used. 'Contribution' refers to the possible reasons for robustness.

Study	Main Author	Components	Contribution
S1	Reddy et al., 2020	Non-uniform and Multiscale sampling	High SNR, Scale and translation invariance
S2	Dapello et al., 2020	V1 neuronal features	Feature extraction and generalization
S3	Qin et al., 2019	Cortical minicolumn and Feedback	Object based representation
S4	Huang et al., 2020	Feedback (self consistency)	Predictive coding
S5	Kim et al., 2020	Anatomical features of visual stream and Feedback	Decorrelation and sparse coding
S6	Li et al., 2019	Neurophysiological data	Inductive bias
S7	Tadros et al., 2019	SNN + STDP	Feature abstraction

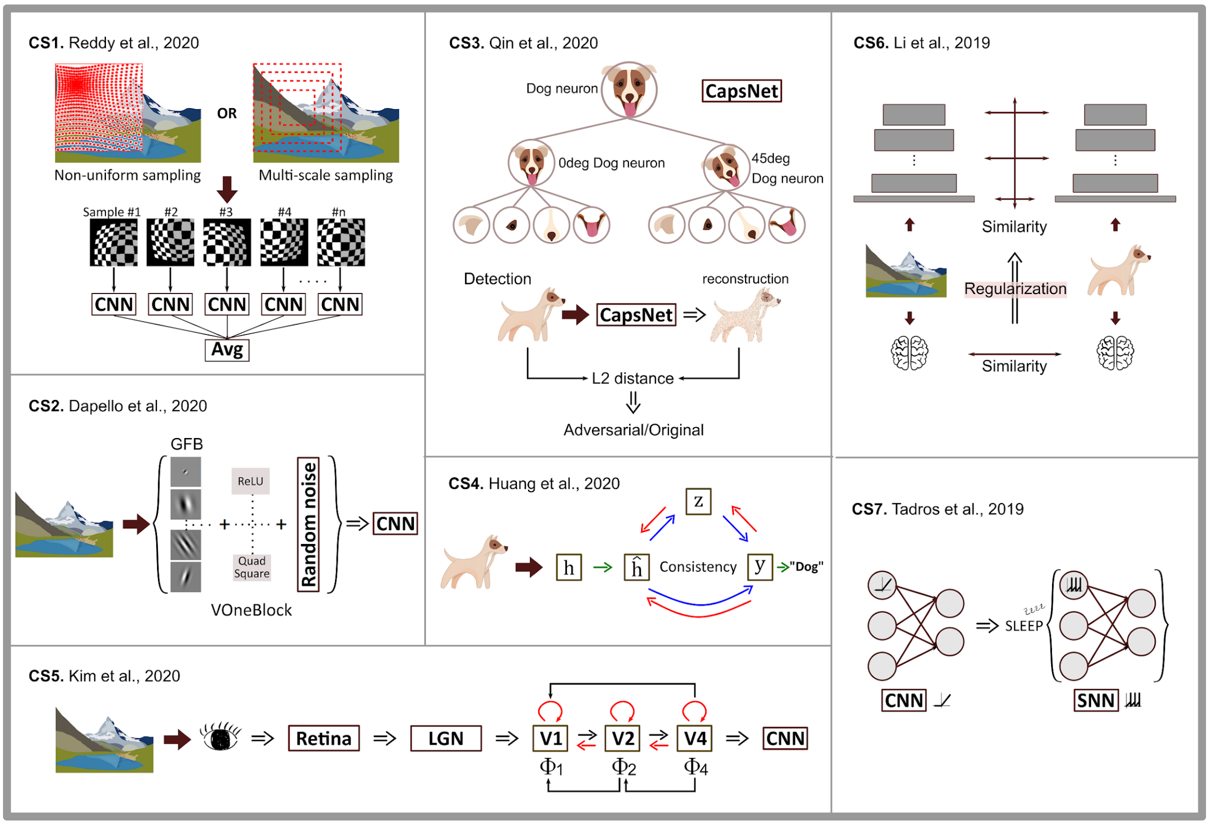


Figure 3. Biologically inspired components. ***S1.*** *Retinal non-uniform and multi-scaling sampling methods implemented by Reddy et al., 2020. Samples were sampled from an image using either non-uniform or multi-scale methods. A DCNN processes each sample and then averages the outputs to obtain the final classification prediction.* ***S2.*** *V1-like block, VOneBlock, prefixed on DCNN proposed by Dapello et al., 2020. VOneBlock consists of a Gaussian Filter Bank (GFB), linear and non-linear activation functions, and stochasticity. It pre-processes the input image and feeds the output into a DCNN.* ***S3.*** *Capsule Network (CapsNet) and the adversarial noise detection method suggested by Qin et al. (2020). Qin et al. (2020) detected the adversarial sample by using the* L_2 *(Euclidean) distance between the input image and the reconstructed image from CapsNet.* ***S4****. Recurrent model proposed by Huang et al., 2020. Input images* h*, are processed by the feedforward DCNN to produce intermediate representations* z *and predict the output label* y*. In the feedback pass, the predicted labels* y *are used to reconstruct* z *and then* h*. The reconstruction (*$\hat{h}$*) of the input image (h), the intermediate latent representation (z), and predicted output label (y) are dynamically modulating each other through feedforward and feedback processes.* ***S5.*** *Overall visual components implemented in the model proposed by Kim et al., 2020. Retina to V4 processes are implemented as a pre-processing module. The input images are first pre-processed by the module and then fed into the DCNN.* ***S6****. Mice V1 activity based DCNN regularization implemented by Li et al., 2019. The authors measured mouse V1 neuronal activation similarities given two different images and then used this similarity measure to regulate the activation patterns of DCNNs.* ***S7****. Sleep model implemented by Tadros et al., 2019. DCNNs are first trained in a standard way and then transformed to SNN, finally, the weights are consolidated by STDP.*

4. Analysis

We have seen how different components of biological vision can help to obtain more robustness against adversarial attacks. In this section, we analyze the evaluation of robustness as carried out by these studies (Table 1) and offer insights for future evaluation.

4.1. Evaluation

In Table 2, we summarize the list of attacks, metrics, datasets, and norms used for different works. As we can see, it is very difficult to compare across methods and understand the significance of the respective bio-inspired component (Table 1).

4.1.1 Experimental Settings

Evaluation of robustness is dependent on the dataset used to generate adversarial perturbations. The robustness of a DCNN on a more complex and realistic dataset like ImageNet (Deng et al., 2009) usually indicates the robustness of the DCNN on less complex datasets (Shafahi et al., 2020) like CIFAR10 (Krizhevsky, 2009). Datasets like ImageNet have more natural and realistic images, and hence more human vision-like features can be learned (Salman et al., 2020). Thus, evaluation of the robustness of DCNN on more complex datasets is a stronger result than on less complex datasets. S1, 2, and 5 have evaluated the robustness of their models with the more complex ImageNet dataset, while the other studies show their robustness on less complex datasets (Table 2). Furthermore, for evaluating a models' robustness, adversarial perturbations should be created w.r.t. the model. All studies (except S5) indeed attack their proposed models.

Then, the choice of ϵ values has a crucial role in analyzing the performance of the different proposals. In S2, for example, accuracy is hard to evaluate since accuracy was averaged across ϵ values and L_p norms (Table 2). While integrating adversarial accuracy over ϵ values is acceptable (Bouniot et al., 2021), it becomes problematic when averaged over two values $\{\frac{1}{1020}, \frac{1}{255}\}$.

Secondly, as mentioned in Sec 2.1, each $L_p$norm has a very different meaning (Tramèr & Boneh, 2019). Therefore, for a fair comparison with other studies, we obtained the non-averaged adversarial accuracy for the main model from the supplementary material provided by the authors. Further, we re-evaluated the robustness of the base models used in the paper without any averaging and presented them in Fig. 4. We observe that the authors reported an average adversarial accuracy (across ϵ values and L_p norms) of 51.1% for their method and 52.3% for Adversarially trained DCNN, but in actuality with the L_∞ norm, the adversarial accuracy differences (Δacc = adv. acc. of their model - adv. acc. of AT DCNN) were different for different ϵ values (ϵ= $\frac{1}{1020}, \Delta acc = 1.1\%$; ϵ= $\frac{1}{255}; \Delta acc = -25\%$; ϵ= $\frac{4}{255}; \Delta acc = -33.58\%$).

*Table 2. Evaluation ‘ * ’ indicates a customized version of the adversarial attack created by the authors. AA indicates Adversarial accuracy, UDR (undetected rate) refers to the fraction of adversarial samples undetected by the method from all the generated samples, MD Indicates the median noise distortion. All the above metrics are calculated for a given value of ϵ and a given L_p. ‘ ✢ ’ indicates adversarial accuracy that had been averaged over various $L_p$norms andϵvalues as mentioned by the authors of S2.*

Study	Image Dataset	L_p norm	Metric	Attacks used
S1	CIFAR10, ImageNet	*1, 2, ∞*	AA	PGD, FGSM, PGD ADAM, L2 CW, DBA, PGD BPDA
S2	ImageNet-Val	*1, 2, ∞*	✢	PGD+MC
S3	MNIST, F-MNIST, SVHN	*2, ∞*	UDR	PGD, PGD-R*, CW, BIM
S4	F-MNIST, CIFAR10	∞	AA	PGD, PGD*, SPSA
S5	ImageNet-Val	x	AA	PGD
S6	CIFAR10	0, 1, 2, ∞	MD	PGD, DBA, etc following list of attacks by Brendel et al., 2019
S7	Patch, MNIST	2	MD	DeepFool, JSMA, DBA, FGSM

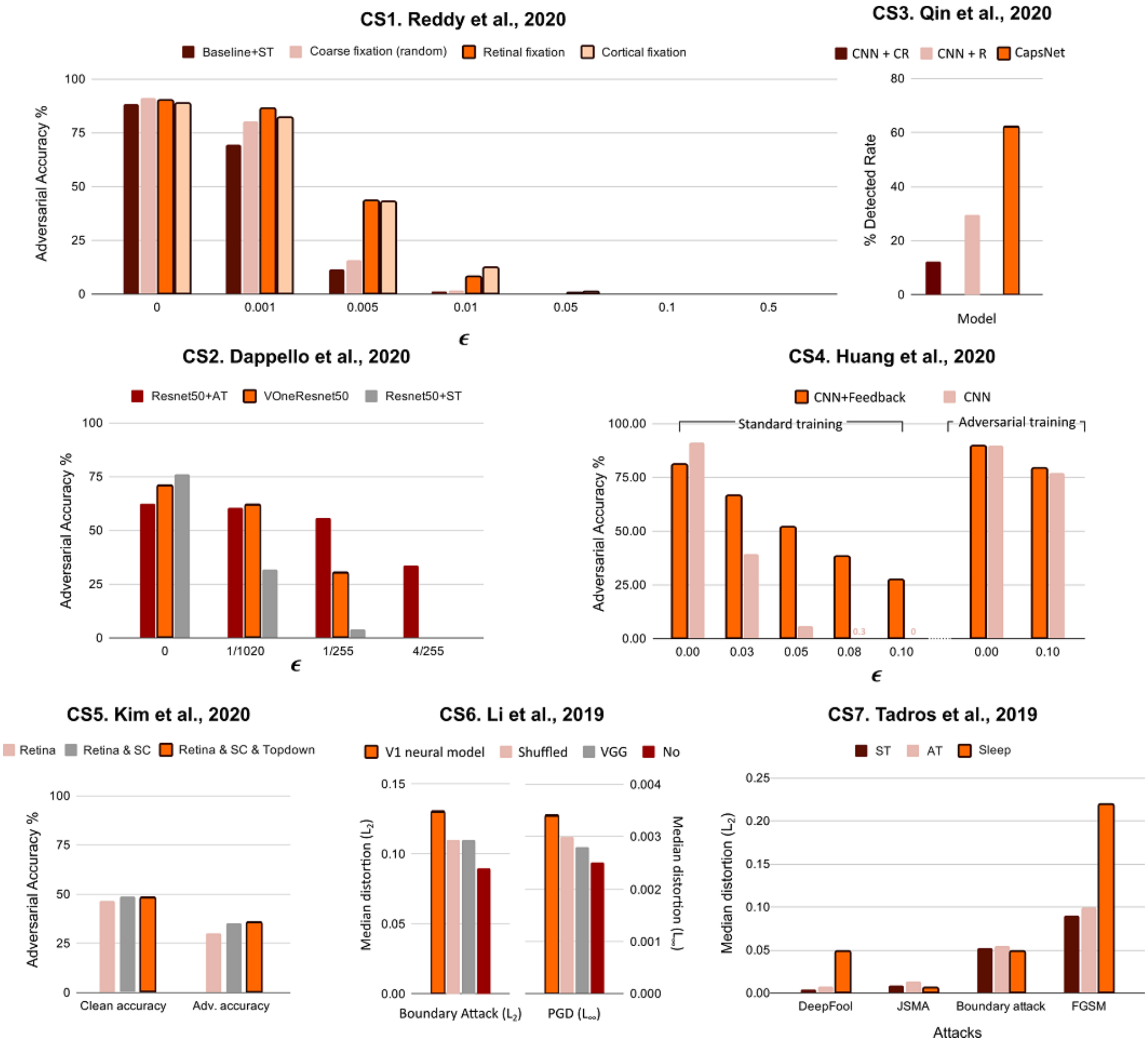


Figure 4. Evaluation summary. The orange bar indicates robustness w.r.t. the metrics used (Table 2) from the proposed contributions, except for S3 where we used (1-UDR) metric for ease of understanding. For all studies, the ***higher*** *the orange bar is, the* ***better*** *is the adversarial robustness of the respective method. From Table 2, we see that many of the studies have worked with multiple datasets and norms. For the sake of easier understanding, we chose to only present the results for the most complex image dataset used (Complexity is approximately: ImageNet (Deng et al., 2009) > CIFAR10 (Krizhevsky, 2009) > SVHN (Netzer et al., 2011) > F-MNIST (Xiao et al., 2017) > MNIST (LeCun & Cortes, 2010)). We also chose to present the results of the common* L_∞ *norm for all works with the exception of S7, which only tested on the* L_2 *norm. Additionally, for S4, we present the results from F-MNIST instead of CIFAR10 since it is the only dataset from the paper for which standard and adversarially trained feedback models are compared. Similarly, for S1, we chose to present the results for ImageNet10, a 10 class subset of ImageNet, since it had a more comprehensive evaluation. For S2 and 5, the ImageNet dataset was used. For S3, we use SVHN; for S6 we use CIFAR10; and for S7, we use MNIST since it was the most complex image dataset used.*

Currently, most of the adversarial attack methods used for evaluation are iterative (Madry et al., 2017; Athalye et al., 2018; Carlini et al., 2019). Madry et al., (2017) showed that with a small increase of the number of iterations (from 7 to 20), adversarial accuracy deteriorates significantly (~ 5%; Table 2 of Madry et al., 2017). This indicates that one should always have *enough iterations* to create the adversarial samples.

A simple way of finding the required number of iterations for an attack is to pit the *number of iterations against adversarial accuracy*. After the required number of iterations, the *adversarial accuracy* will saturate close to 0, indicating that more iterations after this point do not result in better adversarial samples. Only S2 did such an analysis to decide the number of iterations (Table 3). Since this analysis requires large computational resources, it is not easy to do. The alternative is to just give a sufficiently large number of iterations as suggested by Madry et al., (2017) and carried out by S3.

Table 3. Worst-case evaluation parameters: For each study, we picked the most complex image dataset and the strongest attack. All attacks in the table are for L_∞ ***norm****. ' ? ' Indicates values missing in the original paper. We excluded S6 and 7 from the table since the authors used the mean distortion metric to find the minimum perturbation needed to cause a misclassification. For doing so, by definition, a large number of iterations are needed (1000 or more).*

Study	**Dataset**	**Attack**	ϵ ranges	*#iterations*	*#iterations vs Adv. Acc analysis*
S1	ImageNet10	PGD	{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.5}	5 or 20	x
S2	ImageNet-Val	PGD-MC	$\{\frac{1}{1020}, \frac{1}{255}\}$	64	✓
S3	SVHN	PGD-R*	0.1	200	Not required
S4	CIFAR10	PGD*	8/255	7	x
S5	ImageNet-Val	PGD	?	?	?

Further, for S1 and 2, we can see that for the Imagenet dataset (ImageNet10 is a 10 class subset of ImageNet), and the L_∞ norm based attacks, a very different range of ϵ values are used for evaluation. For each dataset and $L_p$norm, the values of ϵ are commonly agreed upon to be imperceptible (Madry et al., 2017; Bouniot et al., 2021). For example for ImageNet, the L_∞norm perturbations are tested with $\epsilon \in \{\frac{4}{255}, \frac{8}{255}, \frac{16}{255}\}$. Note that Elsayed et al. (2018) showed that L_∞ norm perturbations up to $\epsilon \approx \frac{32}{255}$ are imperceptible to humans, which is much larger than the values tested by S2.

4.1.2 Understanding the bio-inspired components

Component analysis (Sec. 3.1). All the studies did the component analysis by freezing each component or alternating the bio-inspired component with a simpler variant (Table 4).

S2 used the Gabor filter parameters sampled from empirical distribution from primate brain (V1 Gabor filter bank), and used the parameters from a uniform distribution (GFB parameters chosen uniformly) as the baseline model. Also, S2 implemented stochasticity using Poisson noise, which is inspired by primate V1 neurophysiology. However, it remains unclear whether bio-inspired noises are any better than random noise. It might be interesting to replace the bio-inspired Poisson noise in the S2 model with random noise to see whether the Poisson noise, determined from empirical data, really matters (Dhillon et al., 2018; Fawzi et al., 2016; Rakin et al., 2018)

S6 used a pairwise similarity matrix based on V1 neuronal activations (Neural similarity) as a regularizer. Thus, S6 was tested against two baselines: a random shuffle of their proposed similarity matrix (Shuffled similarity) and a VGG based similarity matrix (VGG similarity).

Validate components. For S3, 4, 5, and 7, since these studies mainly propose concepts like feedback and sleep for improving robustness, it is hard to find a simpler version of their model to be compared with. The baseline for such studies is the DCNN itself and all seven studies made this comparison.

S3 and 4 visualized the reconstructions using their feedback processes (Table 4). S4 also used Grad-CAM (Selvaraju et al., 2020) visualizations and showed that with more iterations of feedback, their model learned to extract better features from the images. S7 verified memory consolidation by weight visualization of the output layer weights, showing that indeed the connections with weak weights were pruned, and those with stronger weights were strengthened by their method.

Table 4. Summary of the component analysis. Bio-inspired components of each study are listed in the 'Components' column. Under the 'Compared with' column, we list the baseline comparison models, which can either be the full model removing the corresponding component (ablated model) or a simplified model. Under

the 'Visualization' column, we listed the visualization method if the functionalities of the component were visualized.

Study	Components	Compared with	Visualization
S1	Non-uniform Samp.	Uniform Sampling	-
	Multi-scale samp.	Single scale sampling	-
S2	V1 Gabor filter bank	GFB parameters chosen uniformly	-
	Simple cell	Removal	-
	Complex cell	Removal	-
	Stochasticity	Removal & Random noise	-
S3	Feedback	Removal: baseline DCNN	Reconstruction
S4	Feedback	Removal: baseline DCNN	Reconstruction, Grad-CAM
S5	Retinal structure	Removal: baseline DCNN	-
	Sparse coding	Removal	-
	Feedback	Removal	Reconstruction, weight visualization
S6	Neural similarity	Shuffled and VGG similarities	-
S7	SNN	Removal: baseline DCNN	Weight visualization

Attacks for Feedback. Feedback based models usually relay the feedback output from each iteration to the feedforward sweep of the next iteration. These types of models, when unrolled across iterations, are equivalent to extremely deep feedforward models (Kreiman & Serre, 2020). Athalye et al., 2018 have shown that for these deep feedforward models, there is a chance that gradients vanish or explode while calculating the adversarial perturbation (Eq. 1). This is because existing adversarial attack algorithms are not optimized for such deep architectures (taking too many iterations). The authors of both S3 and 4 designed their own variants of the adversarial attacks to account for the vanishing/exploding gradient effect. The robustness evaluation results in S5 are unfortunately incomplete from that point of view.

4.1.3 Evaluation Completeness

Check for gradient masking. As stated in Sec.2.2.2, gradient masking gives a false sense of robustness. As seen in Table 5, most of the studies checked for gradient masking specially the ones which have potential components that could cause gradient masking.

Table 5. Checklist for gradient masking and relevant attacks. 'Components' refers to the feature that possibly causes gradient masking. The numbering in the 'Checklist' column corresponds to the checklist for gradient masking provided in Section 2.2.2. 'Relevant attacks' column refers to similar adversarial attacks, which were used to break gradient masking based defenses having similar components which have been listed in the column, titled 'Similar defenses'. 'Compared to AT' column refers to the studies which have compared themselves to Adversarial Training. For all columns, untick ' ✓ ' represents the studies that did respective columns' analysis and ' x ' represents the opposite; ' - ' represents not relevant to the study. Since all studies, except S3, change decision boundaries by training hence need to be compared to Adversarial Training (AT).

Study	Components	Checklist	Relevant attacks	Similar defense	Compared to AT
S1	Downsampling	1, 2, 4, 5	PGD+BPDA	Guo et al., 2018	✓Only for CIFAR10 but trained on different ϵ
S2	Stochasticity	2	PGD-MC	Dhillon et al., 2018	✓ but trained on different ϵ
S3	-	4	-	-	-
S4	-	2, 3	-	-	✓
S5	Stochasticity	x	x	Dhillon et al., 2018	x
S6	-	6	-	-	x
S7	-	2, 6	-	-	✓

As seen in Table 5, S2 acknowledges that stochasticity based defenses (Dhillon et al., 2018) have been broken in the past (Carlini et al., 2019). By finding the averaged value of the gradient for each attack over many random trials, the stochastic nature of the defense can be broken (PGD-MC; Athalye et al., 2018). Similarly, S1 recognizes that defenses with non-differentiable operations like downsampling and Gaussian blur (Guo et al., 2018) have been broken by utilizing a differentiable approximation of these operations (PGD+BPDA; Athalye et al., 2018). Both S1 and S2 showed that their models were extremely robust even when tested with the PGD-MC and PGD-BPDA attacks, respectively.

Comparison to Adversarial Training. Finally, for defenses, which aim to make DCNNs robust to adversarial perturbations by modifying their decision boundary, the gold standard is to compare their robustness with an Adversarially Trained DCNN. Except for S6 and 7, this was true for all studies (Table 5; Fig. 4).

Further, Adversarial Training depends on the attack used to generate the adversarial samples, which depend on the ϵ parameter and the $L_p$norm. The robustness of Adversarial training depends heavily on the ϵ value (Madry et al., 2017). For a fair comparison of methods, DCNNs should, hence, be adversarially trained and tested on the same ϵ values. However, S1 and 2 compare their model with an adversarially trained DCNN with a fixed ϵ_{train}value whereas the testing was carried out with much lower ϵ_{test}values (S1 used ϵ_{train}= 8/255 and ϵ_{test}: {0.001, 0.005, 0.01, 0.05, 0.1, 0.5}; S2 used ϵ_{train}= 4/255 and ϵ_{test}: {$\frac{1}{1020}, \frac{1}{255}$}).

4.2. Modeling Insights

Many models have proposed that adding biological components can improve robustness to adversarial attacks. However, as we have seen in the last section, whether these attempts were successful cannot easily be judged because often evaluation is hard to carry out. Here, we provide some useful insight gained from these studies and insights for future research directions.

Signal-to-noise ratio (SNR). It seems that S1 and S5 improve robustness by increasing the SNR, which is definitely a viable method. We suggest that this is the case because a high SNR may increase the relevant signals and discard the irrelevant signals (similar to Xie et al., 2019).

Stochasticity. S2 and 5 used stochasticity to improve generalization hence increasing robustness. While in biological systems, all neurons are noisy, S2 and 5 implement stochasticity in only one layer of the DCNN. It would be interesting to see how multiple layers of stochasticity would affect the robustness of the DCNN (Rakin et al., 2018).

Feedback process for the inference and error correction. As seen in Figure 4, S3 and 4 are extremely robust to adversarial perturbations by using feedback connections. Furthermore, with multiple runs of feedforward and feedback processing, it would be possible to learn more abstract and global features capable of generalizing, as shown by Doerig, Bornet, et al., 2020; Doerig, Schmittwilken, et al., 2020 and Kreiman & Serre, 2020, which should in turn increase robustness to adversarial attacks.

Currently, the only limiting factor for running multiple feedback loops in larger models is the large computational requirement, mainly because recurrent models are hard to parallelize for current GPUs. In the future, it may be possible to implement more recurrent long-short range connections like S4 along with inhibitory and excitatory connections as it is done in S3 (Zhao & Huang, 2019).

Other Mechanisms. In S7, a sleep mechanism was implemented using SNN and STDP (memory consolidation). S7 provides good evidence that a well-established cognitive-behavioral component can help to increase robustness. Therefore, we expect that further cognitive features can increase robustness.

S6 showed the possibility of using neurophysiological data for robustness. S6 visualized its proposed similarity matrix to show the V1 neuronal similarity between different image pairs. However, the authors did this only for images without any adversarial noise. It would be interesting if the same similarity matrix could be obtained for a small validation set of adversarial samples to show that the current approach of S6 is already capable of having similar neural activations for adversarial images (like humans).

In summary, the features that improve adversarial robustness are as follows: 1) Increasing the signal-to-noise ratio, as in S1, 2, and 5. 2) Generalization by using stochasticity, as in S2 and 5. 3) Force the network to learn better representations, as in S3, 4, 6, and 7.

5. Discussion

Adversarial Attacks pose a serious problem for state-of-the-art DCNNs in general. In particular, if there are no easy fixes found, it is clear that DCNNs cannot be good models for the human visual system. For this reason, many researchers have tried to defend DCNNs against Adversarial Attacks by using inspirations from the human visual system. Here, we have reviewed the most important recent approaches and found that indeed some bio-inspired components that reproduce stochasticity and feedback phenomena increase robustness. However, due to the immense diversity of evaluation metrics, parameters, or datasets chosen for performance evaluation, it is very hard to judge if a certain component of biological vision is actually useful for increasing robustness (Sec. 4.2). We have reviewed evaluation criteria that could help standardize the evaluation for bio-inspired components. Finally, we summarized insights for future research directions towards designing more robust DCNNs, and generally better understanding of human vision and DCNNs in the presence of perturbations of data.

While we believe such an approach could help the robustness study of DCNN, a question remains.

One man's trash, another man's treasure? Humans primarily use global rather than local visual information. Hence, it may be attractive to encourage the use of more global information for DCNNs through bio-inspired components. However, it is an open question whether including global information can lead to side effects, such as susceptibility to illusions (Watanabe et al., 2018; Pang et al., 2021; Lonnqvist et al., 2021), which, if true, would lead to an explanation of why human vision is often non-veridical.

Through our paper, we have seen how components from the human visual system can help DCNNs to become more robust to adversarial perturbations. However, it is also possible to gain insights from DCNNs to explain phenomena of the human visual system, similar to reinforcement learning (Hassabis et al., 2017). We believe that studying the high vulnerability of current DCNNs can help to explain the adversarial robustness of humans. For example, both Dhillon et al., 2018 and Rakin et al., 2018 showed that random noise increases the robustness of DCNNs. They proposed that this may be due to the generalization properties induced by random noise. This, in turn, tells us that neuronal stochasticity may make humans robust to adversarial attacks. We suggest that such inferences made from the adversarial robustness of DCNNs may, in fact, give us more understanding of how different components of the human visual system function.

In summary, current DCCNs are not good models of the human visual system because they are vulnerable to adversarial attacks. As we have shown, there is not yet a single, simple fix to this vulnerability, as, for example, proposed by Firestone (2019). However, adding many features of the human visual system into DCNNs, in particular feedback and noise, clearly increases the robustness of DCNNs and thus helping bridge the gap between DCNNs and human visual processing. We further argue that we can learn important issues about human visual processing by studying why DCCNs fail to be robust, i.e., we may learn from imperfect models as much as having a perfect model (Lonnqvist et al., 2021). However, commonly agreed criteria on how to evaluate methods to increase robustness are crucial for all these efforts. Our review is a contributive step towards this goal.

Acknowledgement

HM is financed by an interdisciplinary EPFL i-SV grant between Profs. Frossard and Herzog. OHC is financed by the Swiss National Science Foundation (SNF) 320030_176153 “Basics of visual processing: from elements to figures”.

References

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227. https://doi.org/10.1016/j.conb.2017.08.010

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, *6*, 14410–14430.

Athalye, A., Carlini, N., & Wagner, D. (2018a). Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *ArXiv:1802.00420 [Cs]*. http://arxiv.org/abs/1802.00420

Athalye, A., Carlini, N., & Wagner, D. A. (2018b). Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *CoRR*, *abs/1802.00420*. http://arxiv.org/abs/1802.00420

Bakshi, A., & Ghosh, K. (2017). Chapter 26—A Neural Model of Attention and Feedback for Computing Perceived Brightness in Vision. In P. Samui, S. Sekhar, & V. E. Balas (Eds.), *Handbook of Neural Computation* (pp. 487–513). Academic Press. https://doi.org/10.1016/B978-0-12-811318-9.00026-0

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *ArXiv:2004.10934 [Cs, Eess]*. http://arxiv.org/abs/2004.10934

Bouniot, Q., Audigier, R., & Loesch, A. (2021). Optimal Transport as a Defense Against Adversarial Attacks. *ArXiv:2102.03156 [Cs, Stat]*. http://arxiv.org/abs/2102.03156

Brendel, W., & Bethge, M. (2017). Comment on “Biologically inspired protection of deep networks from adversarial attacks.” *ArXiv:1704.01547 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1704.01547

Brendel, W., Rauber, J., & Bethge, M. (2018). Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *ArXiv:1712.04248 [Cs, Stat]*. http://arxiv.org/abs/1712.04248

Brendel, W., Rauber, J., Kümmerer, M., Ustyuzhaninov, I., & Bethge, M. (2019). Accurate, reliable and fast robustness evaluation. *ArXiv:1907.01003 [Cs, Stat]*. http://arxiv.org/abs/1907.01003

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(21), 8621–8644.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., & Kurakin, A. (2019). On Evaluating Adversarial Robustness. *ArXiv:1902.06705 [Cs, Stat]*. http://arxiv.org/abs/1902.06705

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial Attacks and Defences: A Survey. *ArXiv:1810.00069 [Cs, Stat]*. http://arxiv.org/abs/1810.00069

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Choksi, B., Mozafari, M., O'May, C. B., Ador, B., Alamia, A., & VanRullen, R. (2020, October 9). *Brain-inspired predictive coding dynamics improve the robustness of deep neural networks*. NeurIPS 2020 Workshop SVRHM. https://openreview.net/forum?id=q1o2mWaOssG

Croce, F., Andriushchenko, M., Sehwag, V., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2020). RobustBench: A standardized adversarial robustness benchmark. *ArXiv:2010.09670 [Cs, Stat]*. http://arxiv.org/abs/2010.09670

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). *Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations* [Preprint]. Neuroscience. https://doi.org/10.1101/2020.06.16.154542

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 545–559. https://doi.org/10.1016/0042-6989(82)90113-4

De Valois, R. L., William Yund, E., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 531–544. https://doi.org/10.1016/0042-6989(82)90112-2

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J. D., Kossaifi, J., Khanna, A., & Anandkumar, A. (2018, February 15). *Stochastic Activation Pruning for Robust Adversarial Defense*. International Conference on Learning Representations. https://openreview.net/forum?id=H1uR4GZRZ

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S., & Pfeiffer, M. (2015). Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2015.7280696

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., 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.

Echeveste, R., & Lengyel, M. (2018). The redemption of noise: Inference with neural populations. *Trends in Neurosciences*, *41*(11), 767–770.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in Neural Information Processing Systems*, 3910–3920.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., & Madry, A. (2019). Adversarial Robustness as a Prior for Learned Representations. *ArXiv:1906.00945 [Cs, Stat]*. http://

arxiv.org/abs/1906.00945

Fawzi, A., Moosavi-Dezfooli, S.-M., & Frossard, P. (2016). Robustness of classifiers: From adversarial to random noise. *Advances in Neural Information Processing Systems*, 1632–1640.

Fezza, S. A., Bakhti, Y., Hamidouche, W., & Déforges, O. (2019). Perceptual Evaluation of Adversarial Attacks for CNN-based Image Classification. *ArXiv:1906.00204 [Cs, Eess, Stat]*. http://arxiv.org/abs/1906.00204

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *ArXiv:2006.16736 [Cs, q-Bio]*. http://arxiv.org/abs/2006.16736

González, O. C., Sokolov, Y., Krishnan, G. P., Delanois, J. E., & Bazhenov, M. (2020). Can sleep protect memories from catastrophic forgetting? *Elife*, *9*, e51005.

Guo, C., Rana, M., Cisse, M., & Maaten, L. van der. (2018, February 15). *Countering Adversarial Images using Input Transformations*. International Conference on Learning Representations. https://openreview.net/forum?id=SyJ7ClWCb

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. F. (1981). A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, 683–685.

Hosseini, H., Kannan, S., & Poovendran, R. (2019). Dropping Pixels for Adversarial Robustness. *ArXiv:1905.00180 [Cs, Stat]*. http://arxiv.org/abs/1905.00180

Huang, Y., Gornet, J., Dai, S., Yu, Z., Nguyen, T., Tsao, D., & Anandkumar, A. (2020). Neural Networks with Recurrent Generative Feedback. *Advances in Neural Information Processing Systems*, *33*.

Hubel, D. H., & Wiesel, T. N. (1963). Shape and arrangement of columns in cat's striate cortex. *The Journal of Physiology*, *165*, 559–568. https://doi.org/10.1113/jphysiol.1963.sp007079

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, 125–136.

Jordan, M., Manoj, N., Goel, S., & Dimakis, A. G. (2019). Quantifying Perceptual Distortion of Adversarial Examples. *ArXiv Preprint ArXiv:1902.08265*.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern*

Recognition, 3128–3137.

Kaur, S., Cohen, J., & Lipton, Z. C. (2019). Are Perceptually-Aligned Gradients a General Property of Robust Classifiers? *ArXiv Preprint ArXiv:1910.08640*.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kim, E., Rego, J., Watkins, Y., & Kenyon, G. T. (2020). Modeling Biological Immunity to Adversarial Examples. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4665–4674. https://doi.org/10.1109/CVPR42600.2020.00472

Kiritani, T., & Ono, K. (2020). Recurrent Attention Model with Log-Polar Mapping is Robust against Adversarial Attacks. *ArXiv:2002.05388 [Cs]*. http://arxiv.org/abs/2002.05388

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: Feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, *1464*(1), 222–241. https://doi.org/10.1111/nyas.14320

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *BioRxiv*, 408385. https://doi.org/10.1101/408385

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *ArXiv:1607.02533 [Cs, Stat]*. http://arxiv.org/abs/1607.02533

Kwabena Patrick, M., Felix Adekoya, A., Abra Mighty, A., & Edward, B. Y. (2019). Capsule Networks – A survey. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.09.014

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences*, *15*(8), 343–351. https://doi.org/10.1016/j.tics.2011.06.004

Li, Z., Brendel, W., Walker, E. Y., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F. H., Pitkow, X., & Tolias, A. S. (2019). Learning From Brains How to Regularize Machines. *ArXiv:1911.05072 [Cs, q-Bio]*. http://arxiv.org/abs/1911.05072

Lonnqvist, B., Bornet, A., Choung, O. H., Doerig, A., & Herzog, M. H. (2021). *A comparative biology approach to CNN modeling of vision: A focus on differences, not similarities.*

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *ArXiv Preprint ArXiv:1706.06083*.

Marchisio, A., Nanfa, G., Khalid, F., Hanif, M. A., Martina, M., & Shafique, M. (2020). Is Spiking Secure? A Comparative Study on the Security Vulnerabilities of Spiking and Deep Neural Networks. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207297

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2574–2582.

Nayebi, A., & Ganguli, S. (2017). Biologically inspired protection of deep networks from adversarial attacks. *ArXiv:1703.09202 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1703.09202

Nguyen, T., Ho, N., Patel, A., Anandkumar, A., Jordan, M. I., & Baraniuk, R. G. (2019). A Bayesian Perspective of Convolutional Neural Networks through a Deconvolutional Generative Model. *ArXiv:1811.02657 [Cs, Stat]*. http://arxiv.org/abs/1811.02657

Olshausen, B. A. (2013). 20 Years of Learning About Vision: Questions Answered, Questions Unanswered, and Questions Not Yet Asked. In J. M. Bower (Ed.), *20 Years of Computational Neuroscience* (pp. 243–270). Springer New York. https://doi.org/10.1007/978-1-4614-1424-7_12

Ortiz-Jimenez, G., Modas, A., Moosavi-Dezfooli, S.-M., & Frossard, P. (2020). Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *ArXiv Preprint ArXiv:2010.09624*.

Paiton, D. M., Frye, C. G., Lundquist, S. Y., Bowen, J. D., Zarcone, R., & Olshausen, B. A. (2020). Selectivity and robustness of sparse coding networks. *Journal of Vision*, *20*(12), 10–10. https://doi.org/10.1167/jov.20.12.10

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *ArXiv:2102.01955 [Cs, q-Bio]*. http://arxiv.org/abs/2102.01955

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387.

Pennartz, C. M. A., Dora, S., Muckli, L., & Lorteije, J. A. M. (2019). Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. *Trends in Neurosciences*, *42*(9), 589–603. https://doi.org/10.1016/j.tins.2019.07.005

Poggio, T., Mutch, J., & Isik, L. (2014). Computational role of eccentricity dependent cortical magnification. *ArXiv:1406.1770 [Cs, q-Bio]*. http://arxiv.org/abs/1406.1770

Qin, Y., Frosst, N., Sabour, S., Raffel, C., Cottrell, G., & Hinton, G. (2019, September 25). *Detecting and Diagnosing Adversarial Images with Class-Conditional Capsule Reconstructions*. International Conference on Learning Representations. https://openreview.net/forum?id=Skgy464Kvr

Rakin, A. S., He, Z., & Fan, D. (2018). Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness against Adversarial Attack. *ArXiv:1811.09310 [Cs]*. http://arxiv.org/abs/1811.09310

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews*.

Rathbun, D. L., Warland, D. K., & Usrey, W. M. (2010). Spike timing and information transmission at retinogeniculate synapses. *Journal of Neuroscience*, *30*(41), 13558–13566.

Reddy, M. V., Banburski, A., Pant, N., & Poggio, T. (2020, June 29). Biologically Inspired Mechanisms for Adversarial Robustness. *ArXiv:2006.16427 [Cs, Stat]*. Conference on Neural Information Processing Systems. http://arxiv.org/abs/2006.16427

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, *88*(1), 455–463. https://doi.org/10.1152/jn.2002.88.1.455

Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., & Brendel,

W. (2020). A simple way to make neural networks robust against diverse image corruptions. *ArXiv:2001.06057 [Cs, Stat]*. http://arxiv.org/abs/2001.06057

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic Routing Between Capsules. *ArXiv:1710.09829 [Cs]*. http://arxiv.org/abs/1710.09829

Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., & Madry, A. (2020). Do Adversarially Robust ImageNet Models Transfer Better? *ArXiv:2007.08489 [Cs, Stat]*. http://arxiv.org/abs/2007.08489

Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 1262–1273.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Segal, I. Y., Giladi, C., Gedalin, M., Rucci, M., Ben-Tov, M., Kushinsky, Y., Mokeichev, A., & Segev, R. (2015). Decorrelation of retinal response to natural scenes by fixational eye movements. *Proceedings of the National Academy of Sciences*, *112*(10), 3110–3115. https://doi.org/10.1073/pnas.1412059112

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shafahi, A., Saadatpanah, P., Zhu, C., Ghiasi, A., Studer, C., Jacobs, D., & Goldstein, T. (2020). Adversarially robust transfer learning. *ArXiv:1905.08232 [Cs, Stat]*. http://arxiv.org/abs/1905.08232

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528–1540.

Sincich, L. C., Horton, J. C., & Sharpee, T. O. (2009). Preserving Information in Neural Transmission. *Journal of Neuroscience*, *29*(19), 6207–6216. https://doi.org/10.1523/JNEUROSCI.3701-08.2009

Softky, W. R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience*, *13*(1), 334–350. https://doi.org/10.1523/JNEUROSCI.13-01-00334.1993

Song, C., He, K., Lin, J., Wang, L., & Hopcroft, J. E. (2020). Robust Local Features for Improving the Generalization of Adversarial Training. *International Conference on Learning Representations*. https://openreview.net/forum?id=H1lZJpVFvr

Song, S., Miller, K. D., & Abbott, L. F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, *3*(9), 919–926. https://doi.org/10.1038/78829

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. http://arxiv.org/abs/1312.6199

Tadros, T., Krishnan, G., Ramyaa, R., & Bazhenov, M. (2019, September 25). *Biologically* *inspired sleep algorithm for increased generalization and adversarial robustness in deep neural networks*. International Conference on Learning Representations. https://openreview.net/forum?id=r1xGnA4Kvr

Tramèr, F., & Boneh, D. (2019). Adversarial Training and Robustness for Multiple Perturbations. *ArXiv:1904.13000 [Cs, Stat]*. http://arxiv.org/abs/1904.13000

Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. *ArXiv Preprint ArXiv:2002.08347*.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy. *ArXiv Preprint ArXiv:1805.12152*.

Usrey, W. M., & Alitto, H. J. (2015). Visual Functions of the Thalamus. *Annual Review of Vision Science*, *1*(1), 351–371. https://doi.org/10.1146/annurev-vision-082114-035920

Utrera, F., Kravitz, E., Erichson, N. B., Khanna, R., & Mahoney, M. W. (2020). Adversarially-Trained Deep Nets Transfer Better. *ArXiv:2007.05869 [Cs, Stat]*. http://arxiv.org/abs/2007.05869

Vintch, B., Movshon, J. A., & Simoncelli, E. P. (2015). A Convolutional Subunit Model for Neuronal Responses in Macaque V1. *Journal of Neuroscience*, *35*(44), 14829–14841. https://doi.org/10.1523/JNEUROSCI.2815-13.2015

Wamsley, E. J., Tucker, M. A., Payne, J. D., & Stickgold, R. (2010). A brief nap is beneficial for human route-learning: The role of navigation experience and EEG spectral power. *Learning & Memory*, *17*(7), 332–336.

Wang, X., Hirsch, J. A., & Sommer, F. T. (2010). Recoding of sensory information across the retinothalamic synapse. *Journal of Neuroscience*, *30*(41), 13567–13577.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory Motion Reproduced by Deep Neural Networks Trained for Prediction. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00345

Wei, Y., Krishnan, G. P., Komarov, M., & Bazhenov, M. (2018). Differential roles of sleep spindles and sleep slow oscillations in memory consolidation. *PLoS Computational Biology*, *14*(7), e1006322.

Xie, C., Wu, Y., Maaten, L. v d, Yuille, A. L., & He, K. (2019). Feature Denoising for Improving Adversarial Robustness. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 501–509. https://doi.org/10.1109/CVPR.2019.00059

Xie, Cihang, Tan, M., Gong, B., Wang, J., Yuille, A., & Le, Q. V. (2020). Adversarial Examples Improve Image Recognition. *ArXiv:1911.09665 [Cs]*. http://arxiv.org/abs/1911.09665

Zhang, T., & Zhu, Z. (2019). Interpreting adversarially trained convolutional neural networks. *ArXiv Preprint ArXiv:1905.09797*.

Zhang, Y., Foroosh, H., David, P., & Gong, B. (2019). CAMOU: Learning Physical Vehicle Camouflages to Adversarially Attack Detectors in the Wild. *International Conference on Learning Representations*. https://openreview.net/forum?id=SJgEl3A5tm

Zhao, L., & Huang, L. (2019). Exploring Dynamic Routing As A Pooling Layer. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 738–742. https://doi.org/10.1109/ICCVW.2019.00095

Zoran, D., Chrzanowski, M., Huang, P.-S., Gowal, S., Mott, A., & Kohli, P. (2020). Towards

Robust Image Classification Using Sequential Attention Models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9480–9489. https://doi.org/10.1109/CVPR42600.2020.00950

Appendix C

Cretenoud, A. F., Barakat, A., Milliet, A., **Choung, O. H.**, Bertamini, M., Constantin, C., & Herzog, M. H. (2021). How do visual skills relate to action video game performance?. *Journal of vision*, *21*(7): 10, 1-21.

Journal of Vision (2021) 21(7):10, 1–21

How do visual skills relate to action video game performance?

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

Arthur Barakat
Laboratory of Psychophysics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Laboratory of Behavioral Genetics, Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Logitech Europe S.A., Innovation Park EPFL, Lausanne, Switzerland

Alain Milliet
Logitech Europe S.A., Innovation Park EPFL, Lausanne, Switzerland

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

Marco Bertamini
Department of Psychological Sciences, University of Liverpool, Liverpool, UK
Department of General Psychology, University of Padova, Padova, Italy

Christophe Constantin
Logitech Europe S.A., Innovation Park EPFL, Lausanne, Switzerland

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

It has been claimed that video gamers possess increased perceptual and cognitive skills compared to non-video gamers. Here, we examined to which extent gaming performance in CS:GO (Counter-Strike: Global Offensive) correlates with visual performance. We tested 94 players ranging from beginners to experts with a battery of visual paradigms, such as visual acuity and contrast detection. In addition, we assessed performance in specific gaming skills, such as shooting and tracking, and administered personality traits. All measures together explained about 70% of the variance of the players' rank. In particular, regression models showed that a few visual abilities, such as visual acuity in the periphery and the susceptibility to the Honeycomb illusion, were strongly associated with the players' rank. Although the causality of the effect remains unknown, our results show that high-rank players perform better in certain visual skills compared to low-rank players.

Introduction

Basic visual skills, such as contrast detection and orientation discrimination, are the building blocks for visual processing. It has been suggested that playing video games is associated with better performance in these basic perceptual abilities (for reviews, see

Citation: Cretenoud, A. F., Barakat, A., Milliet, A., Choung, O.-H., Bertamini, M., Constantin, C., & Herzog, M. H. (2021). How do visual skills relate to action video game performance?. *Journal of Vision*, *21*(7):10, 1–21, https://doi.org/10.1167/jov.21.7.10.

https://doi.org/10.1167/jov.21.7.10 Received January 25, 2021; published July 16, 2021 ISSN 1534-7362 Copyright 2021 The Authors

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Downloaded from jov.arvojournals.org on 10/04/2021

Bavelier, Shawn Green, Pouget, & Schrater, 2012; Bediou, Adams, Mayer, Tipton, Shawn Green, & Bavelier, 2018; Boot, Blakely, & Simons, 2011; Chopin, Bediou, & Bavelier, 2019). For example, Hutchinson and Stocks (2013) observed that action video game (i.e., a subset of video games, which rely on physical challenges such as hand-eye coordination and reaction time) players (AVGPs) performed better in a random-dot kinematograms task compared to non-video game players (NVGPs). This suggests that AVGPs are better at global motion detection. Likewise, AVGPs were observed to have improved perceptual speed (Dye, Green, & Bavelier, 2009) in the Test of Variables of Attention compared to NVGPs, whereas the speed-accuracy tradeoff was similar in both groups. Li, Polat, Scalzo, and Bavelier (2010) trained participants with video games to establish the causal effect of action gaming on temporal dynamics and observed reduced backward masking performance (i.e., reduced threshold elevation in a masked contrast detection task) in video game players (VGPs) compared to NVGPs. In addition, VGPs outperformed NVGPs in other perceptual skills, such as multiple object tracking (Green & Bavelier, 2006), task-switching (Shawn Green et al., 2012), spatial resolution (Green & Bavelier, 2007), and contrast sensitivity (Li, Polat, Makous, & Bavelier, 2009). These studies were either intervention studies (e.g., Li et al., 2010), that is, participants were trained with a specific video game, or cross-sectional (e.g., Hutchinson & Stocks, 2013).

Studies have also examined the benefits of playing video games on cognitive abilities (for reviews, see Bavelier et al., 2012; Bediou et al., 2018; Campbell, Toth, Moran, Kowal, & Exton, 2018; Spence & Feng, 2010). For example, VGPs showed enhanced change detection performance compared to NVGPs (Clark, Fleck, & Mitroff, 2011). Kowal, Toth, Exton, & Campbell (2018) tested AVGPs and NVGPs with a Stroop test, which tests inhibition, and a Trail-Making test (TMT), which measures processing speed and task-switching abilities. In both tasks, AVGPs showed faster reaction times compared to NVGPs. However, AVGPs made significantly more errors in the Stroop test compared to NVGPs (no significant difference in the TMT), which indicates that inhibitive abilities may be boosted at the expense of a speed-accuracy tradeoff.

Perceptual learning studies have often reported dramatic improvements in perceptual sensitivity. For example, participants improved performance when trained with a bisection stimulus, i.e., they were able to discriminate smaller offsets after training (e.g., Aberg & Herzog, 2009; Grzeczkowski, Clarke, Francis, Mast, & Herzog, 2017; Grzeczkowski, Cretenoud, Mast, & Herzog, 2019). Video gamers may similarly be exposed to learning effects (see Shawn Green, Li, & Bavelier, 2010), resulting in substantial positive effects in both perceptual and cognitive skills (but see Ferguson, 2007). However, perceptual learning was shown to be specific to the orientation (Ball & Sekuler, 1987; Fahle & Morgan, 1996; Grzeczkowski, Cretenoud, Herzog, & Mast, 2017; Schoups, Vogels, & Orban, 1995; Spang, Grimsen, Herzog, & Fahle, 2010), contrast (Sowden, Rose, & Davies, 2002; Yu, Klein, & Levi, 2004), and motion direction (Ball & Sekuler, 1982; Ball & Sekuler, 1987) of the trained stimulus. Hence, learning does not generalize to untrained stimuli, except when using specific training procedures, such as double training (e.g., Xiao, Zhang, Wang, Klein, Levi, & Yu, 2008). Importantly, many aspects of vision, including video gaming, could strongly benefit from a generalization of perceptual learning (Fahle, 2005).

Most studies so far focused on only one—or very few—task(s), and thus it is unclear whether gaming performance is related to some specific skills or to a common factor. In the latter case, we expect to find strong correlations between gamers' performance in visual tasks. However, this prediction is in contrast with the weak evidence for a unique common factor for vision (e.g., Cappe, Clarke, Mohr, & Herzog, 2014; but see Bosten, Goodbourn, Bargary, Verhallen, Lawrance-Owen, Hogg, & Mollon, 2017; for reviews, see Mollon, Bosten, Peterzell, & Webster, 2017; Tulver, 2019). It seems that visual perception is highly multifactorial. For example, there were only weak correlations between the susceptibility to different illusions, whereas strong correlations exist between different variants of the same illusion, suggesting that there are illusion-specific factors (Cretenoud et al., 2019; Cretenoud, Francis, Herzog, 2020; Cretenoud, Grzeczkowski, Bertamini, & Herzog, 2020; Grzeczkowski et al., 2017). Similarly, there seems to be no unique common factor in eye movements (Bargary, Bosten, Goodbourn, Lawrance-Owen, Hogg, & Mollon, 2017), hue scaling (e.g., Emery, Volbrecht, Peterzell, & Webster, 2017), and contrast perception (Bosten & Mollon, 2010; Peterzell, 2016; Peterzell, Schefrin, Tregear, & Werner, 2000).

The popularity of electronic sports (esports), which are video games played at a competitive – sometimes professional – level, has exploded in the last decades with a growing interest in athletes' performance (e.g., Wagner, 2006). One of the leading esports is Counter-Strike: Global Offensive (CS:GO; e.g., Nazhif Rizani & Iida, 2018), a first-person shooter action video game (see Supplementary Figure S1), in which players are split into two groups, that is, terrorists and counterterrorists. Players are usually matched against other players with similar ranks.

Specific motor and cognitive abilities are required to play these video games. For example, flicking, that is, the motor coordination between the player's move and shooting via the computer mouse, is crucial to eliminate the players of the opposite team in first-person shooter video games. Because some of these aspects rely on

Downloaded from jov.arvojournals.org on 10/04/2021

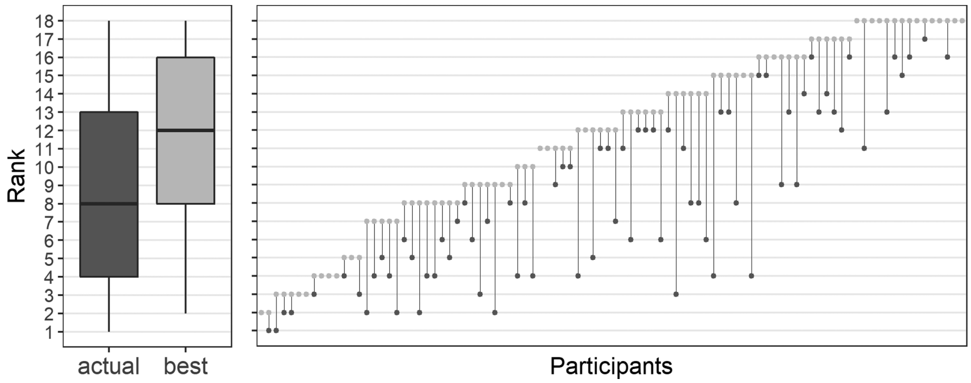


Figure 1. Actual (dark gray) and best (light gray) CS:GO ranks summarized in boxplots (left panel) and shown for each participant (right panel). The higher the rank, the better.

low-level visual skills (e.g., detecting an enemy strongly relies on vision and detection in the periphery), it is of interest to examine different aspects of the game and their relationship with different visual tasks.

To the best of our knowledge, all studies measuring visual abilities in VGPs compared performances between groups, that is, VGPs and NVGPs. However, there is no well-defined criterion to discriminate a VGP from a NVGP. For example, Hutchinson and Stocks (2013) considered participants who played video games for more than 10 hours per week as VGPs, whereas 5 hours per week during the last six months was sufficient in Green and Bavelier (2007). Here, we tested a broad range of CS:GO players, that is, from low- to high-rank players, with a battery of different visual tasks to examine what aspects are associated with expertise, and whether there is a unique, common factor underlying visual perception in AVGPs.

Materials and methods

Participants

Ninety-four participants were recruited (18–35 years; $M = 21.9$; $SD = 3.2$). All participants were AVGPs, played CS:GO at least once in the six months before the experiment, and had a CS:GO rank. Participants signed informed consent prior to the experiment and were paid 20 Swiss Francs per hour. Procedures were conducted in accordance with the Declaration of Helsinki, except for preregistration (§ 35), and were approved by the local ethics committee.

Procedure

The experiment consisted of four parts. First, participants answered a survey about their gaming experience. Participants reported their actual CS:GO rank ($M = 9.0$, $SD = 5.2$; Figure 1), best CS:GO rank ever ($M = 11.8$, $SD = 4.9$; if “best” is not specified, we later refer to the actual ranks; Figure 1), the total ($M = 1232$, $SD = 1879$) and weekly ($M = 13.8$, $SD = 11.7$) number of hours they played CS:GO, and the average number of hours they sleep per night ($M = 7.7$, $SD = 1.1$). Note that ordinal ranks were converted to numerical equivalents from 1 to 18, with 18 being the highest rank (see Supplementary Table S1). Our sample spanned the entire range of ranks, that is, from beginners to experts.

Second, participants performed a battery of 12 visual paradigms: crowding (Crowd), contrast sensitivity (Contrast), the Honeycomb and Extinction illusions (HC/EX), a battery of other illusions (Illusions), N-back (NBack), orientation discrimination (Orient), random dot kinematograms (RDK), simple reaction times (ReacTime), pro- and anti-saccades (Saccade), Freiburg visual acuity (VisAcuity), visual backward masking (VBM), and visual search (VisSrch). The visual paradigms were presented in random order.

Additional variables were extracted for 6 of the 12 paradigms. For instance, the visual search paradigm was tested with two conditions, that is, with either four or 16 distractors. In total, we extracted 38 variables, which are listed in Table 1. When psychometric functions were used (Crowd, Contrast, Orient, RDK, VisAcuity, and VBM paradigms), we discarded blocks when the fit was invalid, i.e., when the point of subjective equality was outside of the search space, the goodness of fit <0.05, or the process did not converge (1.3% of values were discarded).

Third, gaming skills were assessed through six CS:GO mini-games, which were developed by Logitech (Lausanne, Switzerland) in collaboration with the University of Limerick (Ireland) and are publicly available on playmaster.gg. Playmaster is a training space for CS:GO that tests and compares gaming skills among the community and professionals. We extracted

Downloaded from jov.arvojournals.org on 10/04/2021

Paradigm	Variable	Procedure	Lighting condition	Reliability	Feedback	Distance to screen (cm)	Nb of trials	Nb of training trials
Crowd	CrowdSize	PEST	dim	No	Negative	60	96	
	CrowdPeri						160	32
Contrast	Contrast	PEST	dim	No	Negative	200	80	
HC/EX	HC black	Adjustment	on	Yes	No	60	2	1
	HC white						2	
	EX black						2	1
	EX white						2	
Illusions	CS left	Adjustment	on	Yes	No	60	2	1
	CS right						2	
	EB small						2	1
	EB large						2	
	ML in						2	1
	ML out						2	
	PD left						2	1
	PD right						2	
	PZ down						2	1
	PZ up						2	
	TT left						2	1
	TT right						2	
	VH hor						2	1
	VH ver						2	
	WH left						2	1
	WH right						2	
	ZN left						2	1
	ZN right						2	
NBack	NBack		dim	No	Negative	60	40	10
Orient	Orient	PEST	dim	No	Negative	200	80	
RDK	RDK hor	QUEST	on	No	Negative	100	80	4
	RDK rad						80	4
ReacTime	ReacTime		dim	No	No	200	80	
Saccade	proTravel		on	No	Positive and negative	100	16	4
	proSac							
	antiTravel						16	4
	antiSac							
VisAcuity	VisAcuity	QUEST	dim	Yes	No	500	24	24
VBM	VBM	PEST	dim	Yes	Negative	200	80	
VisSrch	VisSrch4		dim	No	Negative	200	40	
	VisSrch16						40	

Table 1. Visual paradigms. Notes: Trials with response times longer than three seconds after the stimulus onset were replaced in the Crowd, Contrast, Orient, RDK, VBM, and VisSrch paradigms. In the Saccade paradigm, positive or negative feedback was provided at the end of each trial as a happy or sad smiley, respectively. In contrast, only negative auditory feedback was provided in the Crowd, Contrast, NBack, Orient, RDK, VBM, and VisSrch paradigms.

six gaming skills: flicking, holding, peeking, shooting, spraying, and tracking (see Supplementary Figure S1), as weighted sums of different features measured in the mini-games.

Fourth and last, participants answered seven self-report questionnaires, which were presented in random order: the Autism-Spectrum Quotient questionnaire (AQ; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), which consists of 50 items; a short version of the Liverpool Inventory of Feelings and Experiences questionnaire (O-LIFE; Mason, Linney, & Claridge, 2005), which investigates positive and negative schizotypy traits with 43 items; the short revised HEXACO personality inventory (HEXACO-60; Ashton & Lee, 2009), which measures 60 items of the six major dimensions of personality (HH:

Downloaded from jov.arvojournals.org on 10/04/2021

honesty-humility; EM: emotionality; EX: extraversion; AG: agreeableness; CO: conscientiousness; OP: openness to experience); the short version of the Barratt Impulsiveness Scale (BIS; Spinella, 2007), which measures impulsivity with 15 items; the Competitiveness Index (CI; Harris & Houston, 2010; Smither & Houston, 1992), which assesses competitive behavior with 14 items measured on a 5-point Likert scale; the Edinburgh handedness inventory (Oldfield, 1971), which assesses participants' hand dominance; and the Personality Research Form dominance subscale (PRFd; Jackson, 1974), which examines social dominance motivation with a 16-item true or false questionnaire. The AQ, BIS, CI, HEXACO, and O-LIFE questionnaires comprise several subscales. Participants could choose between English and French versions of the questionnaires.

Visual tasks and questionnaires were completed at EPFL individually in a quiet room. Because of technical issues, seven participants had to perform the gaming tasks in a gaming room at Logitech (Innovation Park, Switzerland), whereas the others performed the gaming tasks at EPFL. The experimenter stayed in the experimental (or gaming) room with the participant and answered questions at any time.

Apparatus

Stimuli were presented on a BenQ XL2540 LCD monitor (resolution of 1920 × 1080 pixels; screen size: 24.5″) with a refresh rate of 240 Hz. Gaming tasks were performed on an ASUS VG248QE monitor (resolution of 1920 × 1080 pixels; screen size: 24″) with a refresh rate of 144 Hz.

Visual paradigms

Table 1 summarizes details for each visual paradigm, such as the distance to the screen and the light conditions. Stimulus luminance varied between 1 cd/m^2 (black) and 98 cd/m^2 (white).

Crowding

Our paradigm was similar to the one used in Green and Bavelier (2007). First, an E optotype was shown in the periphery, while participants fixated a red dot in the center of the screen (Figure 2a). The red dot was presented for 250 ms. The E optotype was shown for 150 ms with a delay of 100 ms compared to the red dot and at an eccentricity of 10 arcdeg to the right of the red dot. Participants were asked to report the orientation of the optotype within 3 secs by using push buttons, i.e., either standard (right button) or mirrored (left button) orientation. Using an adaptive staircase procedure (parameter estimation by sequential tracking PEST; starting value: 65 arcmin; range value: 10 to 200 arcmin; Taylor & Creelman, 1967), the stimulus size was varied to reach a threshold of 80% of correct responses (CrowdSize).

Second, the task was the same as before and two optotype distractors were added above and below the optotype target. The distractors were randomly oriented in one of the four cardinal directions, and the orientations were counterbalanced in a full factorial fashion. The size of the target was fixed according to the first part of the paradigm (i.e., CrowdSize) and the distance between the target and the distractors was manipulated using a PEST procedure (starting value: 200 arcmin; range value: 6 × CrowdSize to 300 arcmin) to reach 75% of correct responses (CrowdPeri).

Contrast sensitivity

Contrast sensitivity was measured with a 2IFC task (see Lahav, Levkovitch-Verbin, Belkin, Glovinsky, & Polat, 2011). A red fixation dot was presented in the middle of the screen, and subsequently a red and a green circles appeared (2 arcdeg in diameter). Participants indicated in which circle a Gabor patch was presented (spatial frequency: 4.0 cy/arcdeg; duration of presentation: 100 ms; envelope sigma: 0.30 arcdeg; Figure 2b) by pressing a red or green push button, respectively. The mean luminance was 50% and Gabors were rendered using dithering to increase gray level resolution. A PEST procedure (starting value: 10%) was used to measure the contrast threshold level at which participants reached 75% of correct responses.

Honeycomb and extinction illusions

This paradigm was based on a previous study by Bertamini, Herzog, and Bruno (2016; see also Bertamini, Cretenoud, & Herzog, 2019). The Honeycomb and Extinction illusions are characterized by an inability of the participants to see shapes (barbs in the case of the Honeycomb illusion; dots in the case of the Extinction illusion) in the periphery of a uniform texture. The background image (Figures 3a-d) filled the screen. While fixating a red central cross, participants adjusted the size of a red ellipse on the x and y axes using the computer mouse, so that all barbs (Honeycomb) or dots (Extinction) inside the ellipse were perceptible to them.

The red ellipse was displayed with a random size (within the screen size) at the beginning of each trial. Both illusions were tested with two contrast polarity conditions, that is, either black or white barbs in the Honeycomb illusion (Figures 3a-b) and either black or white dots in the Extinction illusion (Figures 3c-d). Hence, there were four conditions (HC black, HC white, EX black, EX white), and each condition was

Downloaded from jov.arvojournals.org on 10/04/2021

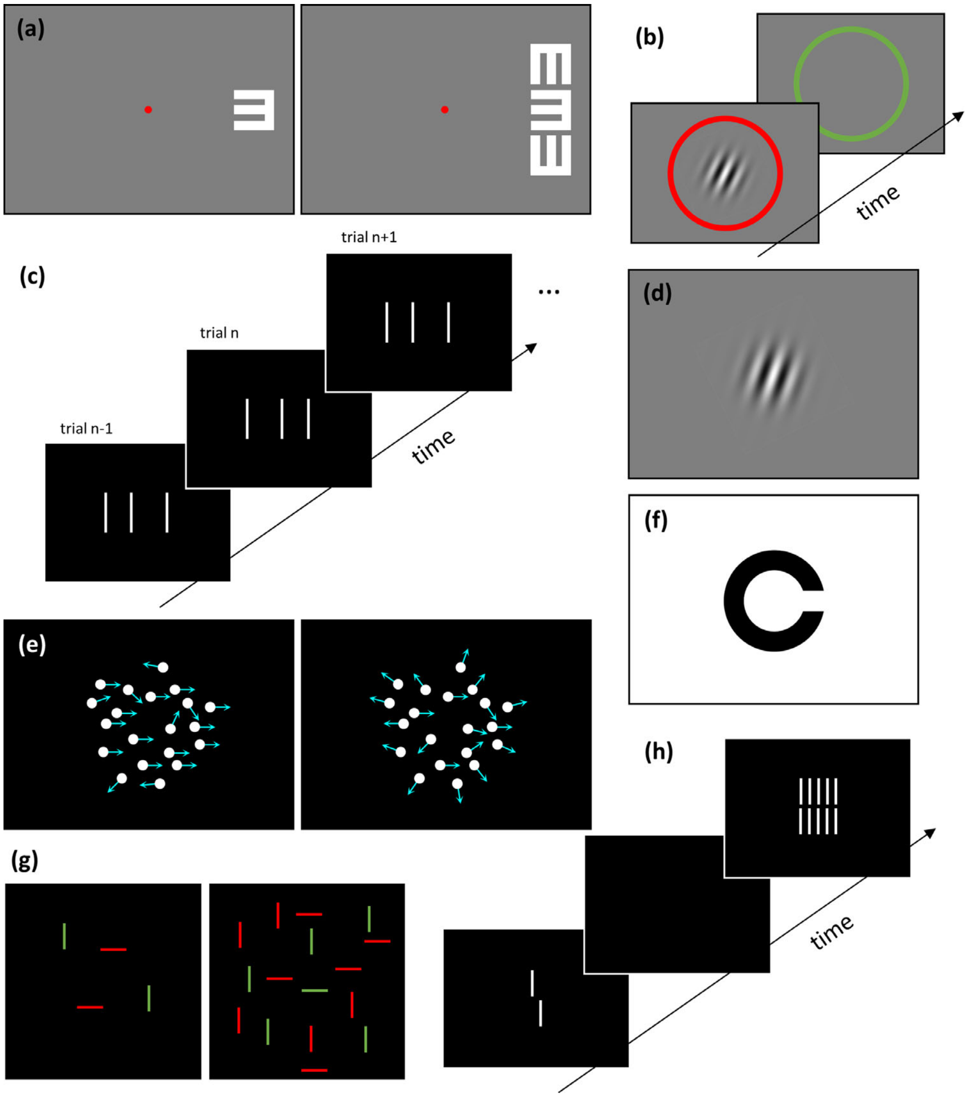


Figure 2. Schematic and exemplary representations of some of the visual paradigms tested. (a) Crowding: the size of the E optotype (left panel) and the distance between the target and distracting optotypes (right panel) varied according to a staircase procedure; (b) contrast sensitivity; (c) N-back with N = 1; (d) orientation discrimination; (e) horizontal (left panel) and radial (right panel) random dot kinematograms (cyan arrows indicate motion direction and were not part of the stimulus); (f) Freiburg visual acuity; (g) visual search (left panel: four-line condition; right panel: 16-line condition); (h) VBM with a five-element grating.

tested twice in a random order. There was no time limit for the adjustment. Random light and dark gray checkerboards (40 random masks presented for 0.5 second each and made of squares of 0.52 arcdeg in side with 0.35 and 0.65 of the maximum luminance) were shown after each trial to reduce the aftereffect. The extracted value was the area of the adjusted ellipse.

Illusions

A battery of nine other illusions was tested (Figures 3e-m): contrast (CS), Ebbinghaus (EB), Müller-Lyer (ML), Poggendorff (PD), Ponzo (PZ), Tilt (TT), vertical-horizontal (VH), White (WH), and Zöllner (ZN). A method of adjustment was used to measure illusion susceptibility, that is, participants were asked to adjust the size (EB, ML, PZ, VH), shade of grey (CS, WH), orientation (TT, ZN), or position (PD) of an element to match the size, shade of grey, orientation, or position, respectively, of a reference on the screen by moving the computer mouse. The reference and adjustable elements were the inside squares in the CS illusion, the central disks in the EB illusion, the vertical segments with inward- and outward-pointing arrows in the ML illusion, the left and right parts of the interrupted diagonal in the PD illusion, the upper and lower horizontal segments in the PZ illusion, the small left and right Gabor patches

Downloaded from jov.arvojournals.org on 10/04/2021

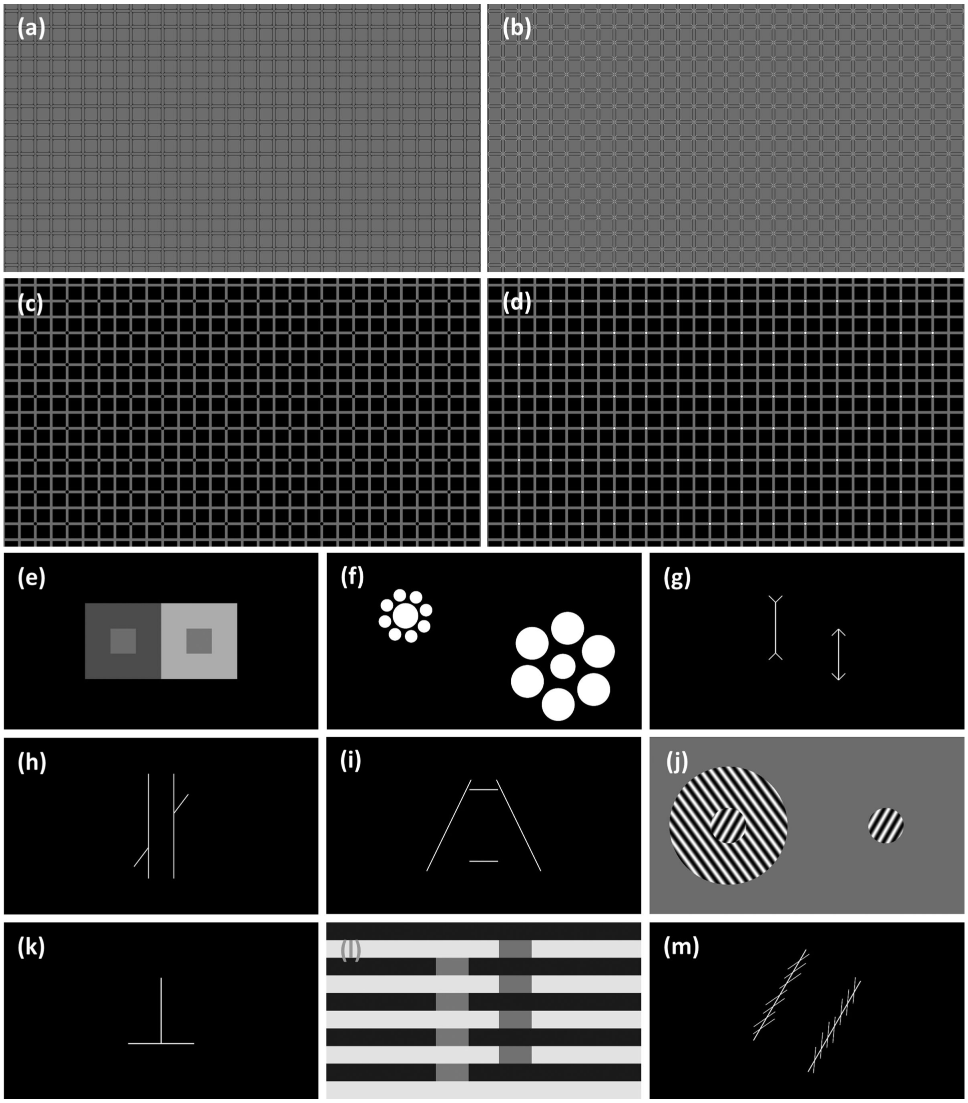


Figure 3. The Honeycomb illusion with (a) black (HC black) and (b) white (HC white) barbs and the Extinction illusion with (c) black (EX black) and (d) white (EX white) dots. The red adjustable ellipse and fixation cross are not depicted here. The images (a) to (d) need to be enlarged so as to fill a large proportion of the visual field; for details, see Bertamini et al. (2016). The battery of other illusions: (e) CS: contrast, (f) EB: Ebbinghaus, (g) ML: Müller-Lyer, (h) PD: Poggendorff, (i) PZ: Ponzo, (j) TT: Tilt, (k) VH: vertical-horizontal, (l) WH: White, and (m) ZN: Zöllner. Illusions (e) to (m) were all tested with two conditions. For example, the upper horizontal line of the Ponzo illusion was adjusted to match the length of the lower horizontal line, or inversely.

in the TT illusion, the horizontal and vertical segments in the VH illusion, the two columns of rectangles in the WH illusion, and the two main streams in the ZN illusion.

Each illusion was tested in two conditions: one element (or series of elements, in the case of the White illusion) was in turn the reference or the adjustable element. For example, in the Ebbinghaus illusion, the task was either to adjust the size of the left central disk so that it appeared to be the same size as the right central disk or to adjust the size of the right central disk so that it appeared to be the same size as the left central disk. The order of presentation of the different illusions and conditions was randomized across participants and there was no time constraint. For a detailed description of the illusions, refer to Cretenoud et al. (2019) and Grzeczkowski et al. (2017). The extracted values were the illusion magnitudes expressed as a difference compared to the reference. Positive and negative illusion magnitudes indicate over- and under-adjustments, respectively.

N-back

We tested a one-back paradigm based on a bisection stimulus, which consists of three vertical lines with the central line being either offset to the left or right compared to the veridical center. The vertical lines were

Downloaded from jov.arvojournals.org on 10/04/2021

1200 arcsec in length and the offset was fixed at 100 arcsec. Each trial consisted in a bisection stimulus, which was shown for 150 ms. Participants were asked to report whether the offset of the current stimulus was on the same or opposite side compared to the offset of the previous stimulus (one-back; Figure 2c) using two push buttons. Forty-one bisection stimuli were shown. We extracted the percentage of correct responses.

Orientation discrimination

Participants performed an adapted version of the orientation discrimination paradigm used in Tibber, Guedes, and Shepherd (2006). Each trial consisted in a red central dot followed by a Gabor patch (spatial frequency: 3.3 cy/arcdeg; duration of presentation: 100 ms; envelope sigma along orientation: 0.57 arcdeg; envelope sigma perpendicular to orientation: 0.19 arcdeg), which was centrally displayed (Figure 2d). Gabors were rendered using dithering to virtually increase gray level resolution. The mean luminance was 50% and the target contrast was 80%. Participants were asked to discriminate between clockwise and counterclockwise stimuli by using two push buttons. The Gabor orientation at which participants gave 75% of correct responses was estimated using a staircase PEST procedure (starting value: 5°).

Random dot kinematograms

The random dot kinematograms paradigm measures global motion perception (Edwards & Badcock, 1995; Hutchinson & Stocks, 2013; Newsome & Park, 1988). Two thousand dots were moving at 5 arcsec/s in a circular aperture (inner diameter: 1 arcdeg; outer diameter: 12 arcdeg) for 500 ms. Each trial consisted of a proportion of dots moving coherently while the rest of the dots moved independently (i.e., distractors; Figure 2e). Participants had to discriminate between leftward and rightward (horizontal, RDK hor) or inward and outward (radial, RDK rad) global motion by using two push buttons. The proportion of dots moving coherently was adapted using a staircase procedure (QUEST with the prior for coherence centered at 60% with SD 50%; Watson & Pelli, 1979; Watson & Pelli, 1983) to reach 75% of correct responses. The two conditions were tested sequentially, and the order was randomized across participants.

Simple reaction times

We used a modified version of the classic Hick-paradigm (Hick, 1952). Participants were instructed to press a mouse button as quickly as possible after a white square (3 arcdeg in side) appeared on a black background. To prevent participants from predicting when the white square appeared, the intertrial interval varied randomly (minimum: 1500 ms; maximum: 3500 ms). The extracted value was the median reaction time (outlier trials were removed using a modified *z*-score; Iglewicz & Hoaglin, 1993).

Prosaccades and antisaccades

Participants gazed at a fixation dot in the center of the screen and were asked to make a prosaccade or an antisaccade toward or away from a target, respectively. The color of the fixation dot, that is, green or red, indicated whether a prosaccade or antisaccade was required, respectively. The target was randomly displayed to the left or to the right of the fixation dot. A positive or negative feedback was provided at the end of each trial as a happy or sad smiley, respectively. Participants were positioned in the head rest of an SMI iViewXHi-Speed 1250 eye tracker (Sensomotoric Instruments, Teltow, Germany), and eye movements were recorded binocularly at 500 Hz. For both prosaccades and antisaccades, we extracted the median travel time (i.e., saccade duration; proTravel and antiTravel), and median saccade time (i.e., delay between the target onset and the saccade onset; proSac and antiSac). As in the simple reaction times paradigm, a modified *z*-score was used to detect and remove outlier trials.

Freiburg visual acuity

Visual acuity was measured following the procedure of the Freiburg visual acuity test (Bach, 1996). Participants were presented with Landolt-C optotypes (Figure 2f) with randomized gap orientations and were asked to indicate the direction of the gap (“up”, “up-right”, “right”, “down-right”, “down”, “down-left”, “left”, or “up-left”) using an eight-button controller. The size of the optotype was varied according to a staircase QUEST procedure, and we extracted the size corresponding to 75% of correct responses.

Visual backward masking

In a visual backward masking paradigm (Herzog, Kopmann, & Brand, 2004; Herzog & Koch, 2001; Roinishvili, Chkonia, Stroux, Brand, & Herzog, et al., 2011), a Vernier stimulus, which consists of two vertical bars offset in the horizontal direction, was presented for 10 ms. The offset between the two horizontal bars was fixed at 75 arcsec. The Vernier stimulus was followed by a variable interstimulus interval, that is, a blank screen, and by a grating for 300 ms (Figure 2h). The grating consisted of five aligned elements of the same length as the Vernier stimulus. Participants were asked to report the offset direction of the lower bar in the Vernier stimulus by using two push buttons. The interstimulus interval vas varied using a PEST procedure (starting

Downloaded from jov.arvojournals.org on 10/04/2021

value: 190 ms) so that participants reached 75% of correct responses.

Visual search

In the visual search paradigm, four (VisSrch4) or 16 (VisSrch16) lines were presented randomly within a black square. Using two push buttons, participants had to report as quickly as possible whether a green horizontal line was present within an array of distractors (green vertical, red vertical, and horizontal lines; Figure 2g). The green horizontal line, that is, the target, was present in 50% of the trials. The median reaction time was extracted for correct trials in both conditions (after outlier trials were excluded according to modified *z*-scores).

Pre-processing and data analysis

Data were extracted in Matlab (MathWorks, Inc., Natick, MA, USA) and analyses were performed in R (R Core Team, 2018), except when mentioned. Alpha level for statistical significance was 0.05.

Reliability

We computed reliability estimates for the variables extracted from visual paradigms, which were tested twice, that is, the Honeycomb and Extinction illusions, the battery of other illusions, visual acuity, and visual backward masking. As suggested by Shrout and Fleiss (1979), two-way mixed effects models (intraclass correlations of type (3,1) or $\mathrm{ICC}_{3,1}$) were computed. Most reliabilities were significant after Bonferroni correction was applied for multiple comparisons (Supplementary Table S2). However, Koo and Li (2016) suggested that ICC coefficients lower than 0.5 are indicative of poor reliability. Hence, variables with 95% confidence interval of the ICC coefficient including 0.5, i.e., the contrast, Ebbinghaus, Müller-Lyer, Ponzo, Tilt, vertical-horizontal, and White illusions, were not considered for further analysis (we excluded both conditions of an illusion even when only one condition showed poor reliability). Note that results were similar when including all variables.

Illusions

Only two illusions showed acceptable reliabilities, namely the Poggendorff and Zöllner illusions. The two conditions of the Poggendorff ($r = 0.721$, $p < 0.001$) and Zöllner ($r = -0.712$, $p < 0.001$) illusions were strongly correlated, suggesting stable individual differences across both conditions. Therefore the two conditions of each illusion were combined into a global illusion magnitude, which was expressed as the sum of the absolute effects in the two conditions.

Bertamini, Cretenoud, and Herzog (2019) recently observed a dissociation between the Honeycomb and Extinction illusions depending on contrast polarity, suggesting that different mechanisms are operating in the black and white conditions of both illusions, respectively. Here, we computed a repeated-measures analysis of variance and similarly observed a significant interaction ($F[1,93] = 118.7$, $p < 0.001$; see Supplementary Figure S2) between the illusion type (Honeycomb or Extinction illusion) and contrast polarity (black or white). Therefore the two conditions of the Honeycomb and Extinction illusions were not combined into a global illusion magnitude.

Data transformation and outlier removal

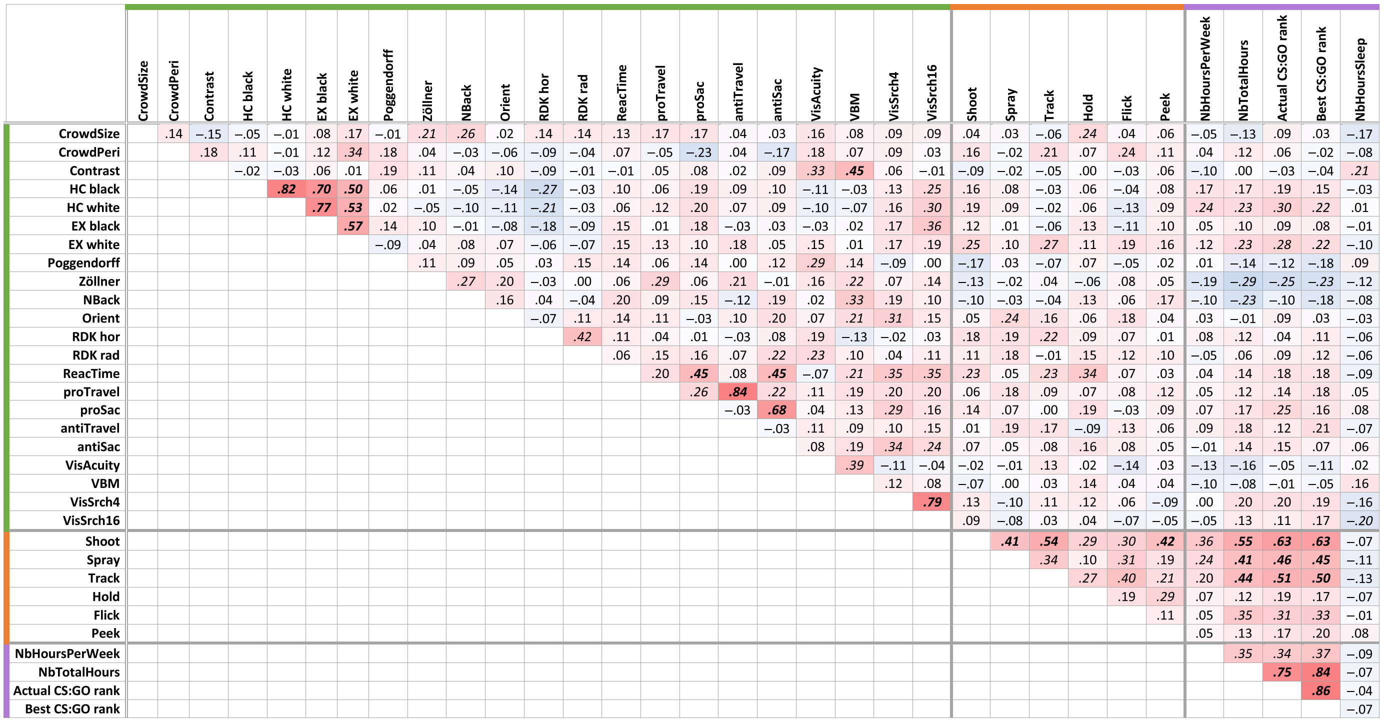
The normality assumption was tested by computing a Shapiro-Wilk test for each variable. Some distributions violated the normality assumption (see Supplementary Table S3). Hence, each distribution was rescaled to approximate a normal distribution. First, we shifted the data distribution to positive values only. Second, we removed outliers based on modified *z*-scores, which are computed from the median and median absolute deviation rather than the mean and standard deviation, respectively, according to a 3.5 criterion (Iglewicz & Hoaglin, 1993). Third, we optimized the λ exponent of a Tukey power transformation (see Supplementary Table S3) to maximize normality according to the Shapiro-Wilk test. Fourth, including the previously removed outliers, data were transformed using the Tukey transformation with the optimized λ parameter. Fifth, we standardized the data by computing modified *z*-scores. Outliers were removed only in the visual variables. Last, we flipped the sign of visual variables when lower values indicated better performance (CrowdSize, CrowdPeri, Contrast, Poggendorff, Zöllner, Orient, RDK hor, RDK rad, ReacTtime, proTravel, proSac, antiTravel, antiSac, VBM, VisSrch4, and VisSrch16). Higher values indicate better performance in all gaming variables.

We imputed outlying and missing values using the “mice” function from the mice R package with method “norm” (Bayesian linear regression with 20 imputation samples) to compute factor analysis and regression models.

Questionnaires

To reduce the complexity of our dataset, the three subscales of the BIS (NI: nonplanning impulsivity, MI: motor impulsivity, AI: attentional impulsivity), which showed strong correlations with each other (NI-MI: $r = 0.441$, $p < 0.001$; NI-AI: $r = 0.406$, $p < 0.001$; MI-AI: $r = 0.532$, $p < 0.001$), were summed in a total score,

Downloaded from jov.arvojournals.org on 10/04/2021



	CrowdSize	CrowdPeri	Contrast	HC black	HC white	EX black	EX white	Poggendorff	Zöllner	NBack	Orient	RDK hor	RDK rad	ReacTime	proTravel	proSac	antiTravel	antiSac	VisAcuity	VBM	VisSrch4	VisSrch16	Shoot	Spray	Track	Hold	Flick	Peek	NbHoursPerWeek	NbTotalHours	Actual CS:GO rank	Best CS:GO rank	NbHoursSleep
CrowdSize		.14	−.15	−.05	−.01	.08	.17	−.01	*.21*	*.26*	.02	.14	.14	.13	.17	.17	.04	.03	.16	.08	.09	.09	.04	.03	−.06	*.24*	.04	.06	−.05	−.13	.09	.03	−.17
CrowdPeri			.18	.11	−.01	.12	*.34*	.18	.04	−.03	−.06	−.09	−.04	.07	−.05	−.23	.04	−.17	.18	.07	.09	.03	.16	−.02	.21	.07	.24	.11	.04	.12	.06	−.02	−.08
Contrast				−.02	−.03	.06	.01	.19	.11	.04	.10	−.09	−.01	−.01	.05	.08	.02	.09	*.33*	***.45***	.06	−.01	−.09	−.02	−.05	.00	−.03	.06	−.10	.00	−.03	−.04	*.21*
HC black					***.82***	***.70***	***.50***	.06	.01	−.05	−.14	*−.27*	−.03	.10	.06	.19	.09	.10	−.11	−.03	.13	*.25*	.16	.08	−.03	.06	−.04	.08	.17	.17	.19	.15	−.03
HC white						***.77***	***.53***	.02	−.05	−.10	−.11	*−.21*	−.03	.06	.12	.20	.07	.09	−.10	−.07	.16	*.30*	.19	.09	−.02	.06	−.13	.09	*.24*	*.23*	*.30*	*.22*	.01
EX black							***.57***	.14	.10	−.01	−.08	−.18	−.09	.15	.01	.18	−.03	.03	−.03	.02	.17	*.36*	.12	.01	−.06	.13	−.11	.10	.05	.10	.09	.08	−.01
EX white								−.09	.04	.08	.07	−.06	−.07	.15	.13	.10	.18	.05	.15	.01	.17	.19	*.25*	.10	*.27*	.11	.19	.16	.12	*.23*	*.28*	*.22*	−.10
Poggendorff									.11	.09	.05	.03	.15	.14	.06	.14	.00	.12	*.29*	.14	−.09	.00	−.17	.03	−.07	.07	−.05	.02	.01	−.14	−.12	−.18	.09
Zöllner										*.27*	.20	−.03	.00	.06	*.29*	.06	.21	−.01	.16	*.22*	.07	.14	−.13	−.02	.04	−.06	.08	.05	−.19	*−.29*	*−.25*	*−.23*	−.12
NBack											.16	.04	−.04	.20	.09	.15	−.12	.19	.02	*.33*	.19	.10	−.10	−.03	−.04	.13	.06	.17	−.10	*−.23*	−.10	−.18	−.08
Orient												−.07	.11	.14	.11	−.03	.10	.20	.07	*.21*	*.31*	.15	.05	*.24*	.16	.06	.18	.04	.03	−.01	.09	.03	−.03
RDK hor													*.42*	.11	.04	.01	−.03	.08	.19	−.13	−.02	.03	.18	.19	*.22*	.09	.07	.01	.08	.12	.04	.11	−.06
RDK rad														.06	.15	.16	.07	*.22*	*.23*	.10	.04	.11	.11	.18	−.01	.15	.12	.10	−.05	.06	.09	.12	−.06
ReacTime															.20	***.45***	.08	***.45***	−.07	*.21*	*.35*	*.35*	*.23*	.05	*.23*	*.34*	.07	.03	.04	.14	.18	.18	−.09
proTravel																*.26*	***.84***	*.22*	.11	.19	.20	.20	.06	.18	.09	.07	.08	.12	.05	.12	.14	.18	.05
proSac																	−.03	***.68***	.04	.13	*.29*	.16	.14	.07	.00	.19	−.03	.09	.07	.17	*.25*	.16	.08
antiTravel																		−.03	.11	.09	.10	.15	.01	.19	.17	−.09	.13	.06	.09	.18	.12	.21	−.07
antiSac																			.08	.19	*.34*	*.24*	.07	.05	.08	.16	.08	.05	−.01	.14	.15	.07	.06
VisAcuity																				*.39*	−.11	−.04	−.02	−.01	.13	.02	−.14	.03	−.13	−.16	−.05	−.11	.02
VBM																					.12	.08	−.07	.00	.03	.14	.04	.04	−.10	−.08	−.01	−.05	.16
VisSrch4																						***.79***	.13	−.10	.11	.12	.06	−.09	.00	.20	.20	.19	−.16
VisSrch16																							.09	−.08	.03	.04	−.07	−.05	−.05	.13	.11	.17	*−.20*
Shoot																								***.41***	***.54***	*.29*	*.30*	***.42***	*.36*	***.55***	***.63***	***.63***	−.07
Spray																									*.34*	.10	*.31*	.19	*.24*	***.41***	***.46***	***.45***	−.11
Track																										*.27*	*.40*	*.21*	.20	***.44***	***.51***	***.50***	−.13
Hold																											.19	*.29*	.07	.12	.19	.17	−.07
Flick																												.11	.05	*.35*	*.31*	*.33*	−.01
Peek																													.05	.13	.17	.20	.08
NbHoursPerWeek																														*.35*	*.34*	*.37*	−.09
NbTotalHours																															***.75***	***.84***	−.07
Actual CS:GO rank																																***.86***	−.04
Best CS:GO rank																																	−.07

Table 2. Correlations between each pair of visual (green), gaming (orange), and CS:GO related (purple) variables expressed as correlation coefficients (Pearson's r). A color scale from blue to red shows the effect sizes from $r = -1$ to $r = 1$. Numbers in italics indicate significant results without correction ($\alpha = 0.05$) and bold numbers indicate significant results with Bonferroni correction ($\alpha = 0.05/990$). See Supplementary Table S4 for the correlations with other questionnaire variables.

which was considered for further analysis. Similarly, we summed the two subscales of the CI (EC: enjoyment of competition, CO: contentiousness) for further analysis, since they significantly correlated ($r = 0.314$, $p = 0.002$). Similarly, we later only considered the total score (i.e., we summed the subscales) of the AQ questionnaire (SS: social skills, AS: attention switch, AD: attention to detail, CO: communication, IM: imagination) and short version of the O-LIFE questionnaire (UE: unusual experiences, IA: introverted anhedonia, CD: cognitive disorganization, IN: impulsive nonconformity). However, the HEXACO personality inventory subscales were kept as separate variables because they showed weak intercorrelations (Table 2).

Results

Correlations

Correlations were computed between each pair of extracted variables (45 variables, 990 comparisons in total) and correlations between visual, gaming, and CS:GO related variables are reported in Table 2. For the sake of readability, correlations with other questionnaire variables (e.g., AQ, BIS, and O-LIFE) are reported in the Supplementary File (Supplementary Table S4). These correlations were weak and mostly nonsignificant ($M_r = -0.008$; $SD_r = 0.119$).

Similarly, correlations were in general weak between pairs of visual variables, except between pairs of variables that were extracted from the same paradigm, for example, between the two conditions of the visual search paradigm (VisSrch4-VisSrch16: $r = 0.790$, $p < 0.001$), and between the Honeycomb and Extinction variables (all $ps < 0.001$; HC black-HC white: $r = 0.819$; HC black-EX black: $r = 0.702$; HC black-EX white: $r = 0.501$; HC white-EX black: $r = 0.767$; HC white-EX white: $r = 0.527$; EX black-EX white: $r = 0.575$), as reported previously (Bertamini et al., 2019). Interestingly, performance in contrast detection and visual backward masking strongly correlated (Contrast-VBM: $r = 0.449$, $p < 0.001$), as previously observed in healthy young adults (da Cruz, Shaqiri, Roinishvili, Favrod, Chkonia, Brand, Figueiredo, & Herzog, 2020).

In contrast, gaming variables showed stronger intercorrelations ($M_r = 0.291$; $SD_r = 0.121$). Similarly, correlations between CS:GO related questionnaire variables (Actual CS:GO rank, Best CS:GO rank, NbHourPerWeek, NbTotalHours) and gaming variables were rather strong ($M_r = 0.314$; $SD_r = 0.182$), which was expected since expertise is gained

Downloaded from jov.arvojournals.org on 10/04/2021

through training (for example, see Macnamara, Hambrick, & Oswald, 2014). For instance, the actual CS:GO rank strongly related to the total number of hours played (Actual CS:GO rank-NbTotalHours: $r = 0.748$, $p < 0.001$) but also to the gaming skills, such as the performance in the shooting mini-game (Actual CS:GO rank-Shoot: $r = 0.629$, $p < 0.001$).

Importantly, the actual CS:GO rank significantly correlated with some visual variables, namely with the Honeycomb white illusion (HC white: $r = 0.298$, $p = 0.004$), Extinction white illusion (EX white: $r = 0.278$, $p = 0.007$), Zöllner illusion ($r = -0.249$, $p = 0.016$), saccade time in pro-saccades (proSac: $r = 0.251$, $p = 0.015$), and with three personality traits (HEXACO HH: $r = 0.232$, $p = 0.024$; HEXACO OP: $r = -0.275$, $p = 0.007$; PRFd: $r = -0.254$, $p = 0.014$). However, not all of these correlations survived Bonferroni correction. Overall, correlations between pairs of visual variables were weak, while we observed stronger correlations between pairs of gaming related variables (including variables related to the rank and amount of training).

Exploratory factor analysis

In order to explore whether a strong and unique factor underlies vision in action video game players and to keep the participant/variable ratio as large as possible, only the visual variables (22 variables) were subjected to an exploratory factor analysis (EFA). The Kaiser-Meyer-Olkin test for sampling adequacy was computed to quantify the degree of intervariable correlations. Visual variables that showed an unacceptable measure of sampling adequacy (i.e., MSA < 0.5) were removed sequentially until all variables showed an acceptable MSA. Four variables were therefore removed for the EFA (Poggendorff, Zöllner, antiTravel, and CrowdSize). The global MSA index after variable removal was 0.659.

Factors were extracted with a common factor analysis to reflect the variance shared between variables (i.e., the common variance). We used an oblique rotation (promax; see Costello & Osborne, 2005) because we had no reason to preclude factors to correlate.

A parallel analysis suggested a five-factor model, whereas only three factors were suggested by scree plot inspection (see Supplementary Figure S3). Because the eigenvalues for factors 4 and 5 were very close to those of a resampled dataset and below 1.0 (Kaiser, 1970; RF1: 3.058; RF2: 1.892; RF3: 1.055; RF4: 0.719; RF5: 0.497), we retained the three-factor model (TLI = 0.615; RMSEA = 0.112 with 90% CI [0.093, 0.134]). The three factors together explained 37.6% of the variance (RF1: 15.6%; RF2: 13.8%; RF3: 8.2%). Loadings are reported in Table 3. According to a simulation published in Hair, Black, Babin, Anderson, and Tatham, (2018), loadings larger than 0.55 are considered as significant with a sample size of 100.

	RF1	RF2	RF3
CrowdPeri	0.144	–0.341	**0.603**
Contrast	0.010	–0.088	**0.586**
HC black	**0.826**	0.147	–0.135
HC white	**0.887**	0.160	–0.161
EX black	**0.842**	0.135	–0.028
EX white	**0.596**	0.082	0.163
NBack	–0.038	0.204	0.249
Orient	–0.109	0.205	0.228
RDK hor	–0.261	0.155	–0.017
RDK rad	–0.129	0.243	–0.028
ReacTime	0.041	**0.646**	–0.071
proTravel	0.085	0.288	0.113
proSac	0.077	**0.660**	–0.130
antiSac	–0.067	**0.799**	–0.096
VisAcuity	–0.068	–0.047	0.507
VBM	–0.018	0.208	**0.586**
VisSrch4	0.203	0.544	0.059
VisSrch16	0.297	0.515	–0.017

Table 3. Rotated factor loadings from an EFA on the visual variables only and after promax (i.e., oblique) rotation. A color scale from blue (negative loadings) to red (positive loadings) is shown. Factor loadings larger than 0.55 are highlighted (bold).

The first factor was mainly related to the Honeycomb and Extinction variables (HC black, HC white, EX black, EX white). Both illusions are related to visual perception in the periphery and were here (Table 2) and previously shown to strongly correlate (Bertamini et al., 2019). The second factor mainly loaded on variables related to reaction times, such as ReacTime, VisSrch4, and VisSrch16, and to the prosaccade and antisaccade paradigm (e.g., proSac, antiSac). The third factor strongly loaded on the contrast detection, visual backward masking (VBM; i.e., a measure of spatiotemporal perception, which may reveal specifically tuned to gaming), and crowding paradigm, which is a measure of visual acuity in the periphery. Interfactor correlations were mostly weak (RF1-RF2: $r = 0.009$; RF1-RF3: $r = -0.011$; RF2-RF3: 0.260). Hence, it seems that there is no strong and unique factor underlying visual perception in action video game players but rather multiple factors, which only explain a small proportion of the variability.

Regression

First, a multiple regression model was computed to estimate how much variance in the players' rank is accounted for by visual performance, gaming skills, and personality traits. Second, we examined the accuracy

Downloaded from jov.arvojournals.org on 10/04/2021

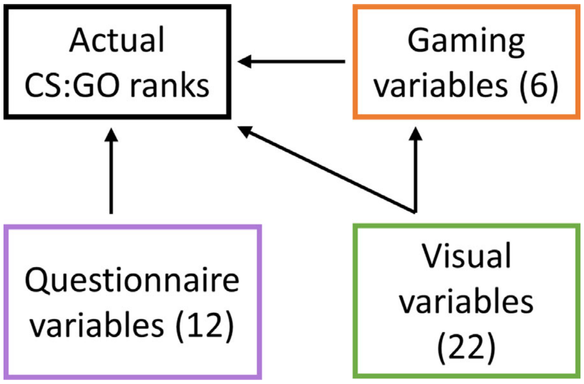


Figure 4. Schematic representation of the path model computed to determine to what extent the players' rank can be predicted by the visual, gaming, and questionnaire variables. The numbers in brackets indicate the number of variables considered. The visual variables not only regressed on the actual CS:GO ranks but also on the gaming variables.

in the prediction of the ranks while reducing the high-dimensionality of the model (i.e., the number of variables). To this aim, we computed an elastic net model, which extracted the variables with stronger predictive power. Importantly, note that the CS:GO related questionnaire variables (i.e., Best CS:GO rank, NbHoursPerWeek, and NbTotalHours) and amount of sleep (NbHoursSleep) were not included for further analysis. Indeed, performance in several domains, such as games, sports, and music, is known to be closely related to the amount of practice (e.g., Macnamara et al., 2014). Here, we aimed at examining whether players' rank can be predicted from variables that are not specifically related to the amount of training, namely visual perception, gaming skills, and personality traits.

Path model

We wondered to what extent the visual, gaming, and questionnaire scores predict the actual CS:GO ranks of the players, and how gaming variables relate to visual variables. Hence, we designed a complex, multiple regression model (i.e., a path model) schematically represented in Figure 4. The actual CS:GO rank is an outcome variable (i.e., endogenous variable), whereas the visual and questionnaire variables are predictors (i.e., exogenous variables). The gaming variables are both outcomes and predictors.

Standardized path coefficients are reported in Table 4 (no correction was applied for multiple comparisons). The visual, gaming, and questionnaire variables explained 69.6% of the variance of the CS:GO ranks. Between 12.9% and 37.4% of the variance of each gaming variable was accounted for by the visual variables. Some gaming variables showed significant standardized path coefficients on the CS:GO ranks, and so did some variables related to visual paradigms (crowding, Honeycomb illusion, Zöllner illusion, random dot kinematograms). Similarly, visual variables showed some significant standardized path coefficients on the gaming variables. For example, the ReacTime variable significantly loaded on the Shoot, Track, and Hold gaming variables.

Hence, it seems that the variance in the players' rank is largely accounted for by performance in visual perception, specific gaming skills, and personality traits.

Elastic net model

Using the scikit-learn package in Python (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, … Duchesnay, 2011), we aimed at predicting the actual CS:GO players' ranks by fitting an elastic net model (Zou & Hastie, 2005), i.e., a regressor, which both uses L1 and L2 regularizations, therefore reducing the dimensionality of the model and the risk of overfitting.

The dataset was split into a training (80%) and test (20%) set. Using a search grid with a fivefold cross-validation, we optimized the model's generalization performance on the training set by tuning two hyperparameters, namely alpha and the L1 ratio (i.e., L1/(L1+L2)). The lower alpha, the more complex the model (i.e., less strict regularization).

Performance on the training set was optimized for alpha = 0.15 and with an L1 ratio of 0.45. With these values for the hyperparameters, the training and test set accuracies were $r^2 = 0.643$ and $r^2 = 0.210$, respectively. The *MSE*s for the training and test sets were 0.21 and 0.26, respectively. In contrast, a dummy regressor resulted in an *MSE* of 0.60 and 0.37 in the training and test sets, respectively. The following variables had nonzero coefficients: CrowdSize (0.034), CrowdPeri (−0.022), HC white (0.088), Zöllner (−0.106), proSac (0.080), VisAcuity (−0.033), VisSrch4 (0.011), Shoot (0.222), Spray (0.065), Track (0.166), Flick (0.005), BIS (0.052), HEXACO CO (−0.008), HEXACO OP (−0.089). Gaming variables were expected to show non-zero coefficients, because they are obviously related to the players' rank (Table 2).

Our results suggest that the Honeycomb illusion and crowding variables are predictors of the players' rank, i.e., players who perceived barbs in larger areas (HC white) and who needed a smaller optotype to achieve 75% of performance (CrowdSize) tend to have higher ranks. Note that both paradigms are related to visual perception in the periphery. However, participants with higher ranks tend to have worse visual acuity in the fovea (VisAcuity) and to be more susceptible to the Zöllner illusion (ZN). In addition, our results suggest that faster reaction times (proSac and VisSrch4) are associated with higher ranks. Lastly, participants with weaker conscientiousness (HEXACO CO), weaker openness to experience (HEXACO OP), and with

Downloaded from jov.arvojournals.org on 10/04/2021

	Actual CS:GO rank	Shoot	Spray	Track	Hold	Flick	Peek
CrowdSize	0.230**	–0.049	0.005	–0.155	0.221*	–0.025	–0.017
CrowdPeri	0.139	0.176	–0.043	–0.139	–0.133	0.127	–0.166
Contrast	0.096	–0.089	–0.073	–0.115	0.004	–0.039	0.074
HC black	0.029	0.060	0.039	–0.057	0.048	0.080	–0.010
HC white	0.297*	0.084	0.163	0.108	–0.048	–0.277	0.005
EX black	–0.157	0.051	–0.043	–0.186	0.175	–0.070	0.050
EX white	–0.192	0.051	0.026	0.367**	0.013	0.390**	0.184
Poggendorff	–0.022	–0.243*	0.078	0.032	0.078	0.055	0.040
Zöllner	–0.303***	–0.118	–0.033	0.123	–0.182	0.130	–0.028
Nback	–0.055	–0.099	–0.032	–0.126	–0.012	–0.063	0.232
Orient	0.067	0.121	0.295**	0.149	0.060	0.050	0.039
RDK hor	–0.212*	0.178	0.268*	0.370***	0.206	0.188	0.021
RDK rad	0.135	0.059	–0.011	–0.256*	–0.090	0.095	0.153
ReacTime	–0.030	0.272*	0.011	0.226*	0.238*	–0.023	–0.048
proTravel	0.124	–0.055	–0.253	–0.162	0.231	–0.052	0.055
proSac	0.167	0.190	0.175	–0.046	–0.050	–0.075	0.065
antiTravel	–0.065	0.088	0.417*	0.255	–0.196	0.118	0.041
antiSac	0.010	–0.140	–0.116	–0.048	–0.054	0.150	–0.088
VisAcuity	–0.081	0.040	–0.139	0.100	–0.037	–0.412***	–0.070
VBM	–0.047	–0.019	0.148	0.078	0.165	0.147	0.036
VisSrch4	–0.015	0.057	–0.088	0.269	0.180	0.127	–0.092
VisSrch16	0.116	–0.119	–0.159	–0.304	–0.271	–0.254	–0.061
Shoot	0.223**						
Spray	0.122						
Track	0.473***						
Hold	–0.095						
Flick	0.072						
Peek	0.029						
AQ	0.088						
BIS	–0.130						
CI	0.128						
HEXACO HH	0.100						
HEXACO EM	–0.117						
HEXACO EX	0.195*						
HEXACO AG	0.009						
HEXACO CO	–0.213*						
HEXACO OP	–0.254**						
Handedness	0.020						
O-LIFE	0.097						
PRFd	–0.160						
r^2	0.696	0.244	0.220	0.374	0.250	0.296	0.129

Table 4. Standardized path coefficients ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$) from the path model (see Figure 4) and variance explained (r^2) of each endogenous variable. The strength of the standardized path coefficients is indicated with a color scale from blue (negative loadings) to red (positive loadings).

Downloaded from jov.arvojournals.org on 10/04/2021

higher score on the Barratt Impulsiveness Scale (BIS), tend to have higher ranks.

Overall, the model drastically reduced the dimensionality of the dataset (from 40 to 14 variables) by extracting the variables with stronger predictive power, such as the Honeycomb illusion and crowding variables.

Discussion

Summary

We tested 94 CS:GO players ranging from beginners to experts with 12 visual paradigms, specific gaming skills, and personality traits to examine what aspects are associated with expertise, and whether there is a unique, common factor underlying visual perception in AVGPs.

First, we observed only weak correlations between visual variables, except between variables that belong to the same paradigm, which can be taken as a measure of reliability. In addition, gaming variables showed strong intercorrelations. A factor analysis revealed three factors explaining about 38% of the variance, which suggests a poor factor structure.

Second, a path model showed that almost 70% of the variance of the actual players' rank is predicted by visual, gaming, and questionnaire scores. Not only gaming variables but also some visual and questionnaire scores showed strong loadings on the players' rank.

Last, we computed an elastic net model to select the features with stronger predictive power on the actual ranks (i.e., to reduce the dimensionality of the dataset). The model retained 14 variables (among which seven were visual variables), which altogether led to better predictions of the ranks compared to a dummy model. The visual variables, which were retained in the elastic net model and showed significant standardized loadings in the path model, were CrowdSize (crowding size), HC white (Honeycomb illusion with white barbs), and the Zöllner illusion. Note that the best CS:GO rank, amount of training (NbHourPerWeek and NbTotalHours), and amount of sleep (NbHoursSleep) were not included in the path and elastic net models.

Importantly, the path model accounted for a larger proportion of the variance in the data compared to the elastic net model ($r^2 = 0.696$ versus $r^2_{training} = 0.643$, respectively), because the former used more variables than the latter (40 versus 14 variables, respectively). While the dimensionality of the dataset was reduced in the elastic net model, the decrease in performance compared to the path model was rather small, which suggests that most variables do not significantly predict the players' rank. However, the test set accuracy was much lower than the training set accuracy in the elastic net model ($r^2_{test} = 0.210$ versus $r^2_{training} = 0.643$), which suggests overfitting even though the elastic net model showed a better test *MSE* compared to a dummy regressor. The small test sample size (20%, i.e., 19 participants only) may partially explain the rather low test set accuracy.

We expected many aspects of gaming to rely on (low-level) visual skills. However, our results suggest that there is no strong common factor for visual perception in CS:GO players. Similarly, there is only weak evidence for a common factor for visual perception in general (Mollon et al., 2017; Tulver, 2019). For example, many specific factors were reported in oculomotor tasks (Bargary et al., 2017), in the perception of faces (Verhallen et al., 2017), and in the susceptibility to visual illusions (e.g., Cretenoud et al., 2019; Grzeczkowski et al., 2017). More generally, basic visual paradigms only weakly correlate with each other (e.g., Cappe et al., 2014).

Positive association between peripheral vision and the players' rank

Rather than a strong common factor for visual perception in CS:GO players, specific visual paradigms and personality traits seem to be strongly predictive of the players' rank. For example, players who perceived more barbs in the Honeycomb white illusion, tended to have higher ranks. Since the four variables extracted from the Honeycomb and Extinction paradigm (HC black, HC white, EX black, and EX white) strongly correlate with each other (Table 2), we expected that either all four variables or none would be significantly associated with the players' rank. However, only one variable (i.e., HC white) showed a non-zero coefficient in the elastic net model (0.088) and a significant standardized path loading (0.297), suggesting that not all variants of the illusions are associated with the players' rank.

Interestingly, similar illusion magnitudes were observed in the Honeycomb illusion with black and white barbs (HC black and HC white; see Supplementary Figure S2), while the mean extent of visible region was previously shown to be larger in the white compared to the black variant (Bertamini et al., 2019). A difference in the experimental design may explain the discrepancies in the results. To estimate the mean extent of the region in which barbs were visible, participants adjusted the size of an ellipse (i.e., on both x and y axes) in the present investigation, whereas a disk (i.e., a single dimension) was adjusted in Bertamini et al. (2019). The background images were the same in both studies. Note that barbs (or disks in the Extinction illusion) were removed during the adjustment in Bertamini et al. (2019), unlike in the present investigation. Despite these differences,

Downloaded from jov.arvojournals.org on 10/04/2021

an interaction between the illusion type (Extinction, Honeycomb) and contrast polarity (black, white) was observed in both studies. It may be worth considering that the magnitude of these illusions conflate a perceptual and a response bias aspect. Participants may differ in the tendency to report what they may "know" rather than what they see, or, even without awareness of this, to “cheat” by not maintaining fixation.

The crowding paradigm similarly seems to be a strong predictor of the players’ rank. Higher ranks were associated with a better visual acuity in the periphery, as reflected by the CrowdSize variable (standardized path coefficient: 0.230; coefficient from the elastic net model: 0.034). Although Green and Bavelier (2007) previously reported an increased spatial resolution in AVGPs compared to NVGPs, we observed a negative association between spatial resolution, as measured with a crowding paradigm (CrowdPeri), and the players’ rank (coefficient from the elastic net model: −0.022). However, further investigation is needed to verify this association, since CrowdPeri did not show up as a significant coefficient in the path model, which may indicate that the association is unreliable. In addition, a radial-tangential anisotropy was reported in crowding (Chung, 2013; Greenwood, Szinte, Sayim, & Cavanagh, 2017), suggesting that the association may be different along the horizontal axis.

Both the Honeycomb illusion and crowding paradigm are related to peripheral vision and were strongly associated with the players’ rank. However, the Honeycomb and crowding variables did not correlate ($M = 0.012$, $SD = 0.067$; Table 2). As in foveal vision, it is likely that vision in the periphery is multifactorial, i.e., there is no strong common factor for peripheral vision. For example, Yashar, Wu, Chen, and Carrasco (2019) reported no common mechanism for crowding across different visual features. The authors tested different visual features under crowding to determine at which processing stage crowding occurs. They observed that orientation and spatial frequency errors were interdependent, whereas orientation and color errors were independent, suggesting that peripheral vision is feature-dependent.

While different features are likely processed differently in the periphery, our results suggest that peripheral vision in general plays an important role in CS:GO. Specifically, it seems that high-rank players have better peripheral vision compared to low-rank players, which adds to previous results reporting evidence for better peripheral vision in AVGPs compared to NVGPs (for a review and meta-analysis, see Chopin et al., 2019). Similarly, increased peripheral visual skills are beneficial to team sports players, such as basketball or soccer players (Faubert & Sidebottom, 2012; Knudson & Kluka, 1997). However, note that the crowding variables only weakly correlated with the players’ rank (CrowdSize: $r = 0.085$, $p = 0.414$; CrowdPeri: $r = 0.057$, $p = 0.656$; Table 2), suggesting that their role in CS:GO is important when interacting with other specific skills only. In contrast, the correlation between the Honeycomb white variable and players’ rank was medium to large ($r = 0.298$, $p = 0.003$), according to Cohen (1988) and Gignac and Szodorai (2016), respectively.

Negative association between central vision and the players’ rank

Surprisingly, visual acuity in the fovea (VisAcuity) was negatively associated with the players’ rank (coefficient from the elastic net model: −0.033). Patino, McKean-Cowdin, Azen, Allison, Choudhury, and Varma (2010) reported that central and peripheral visual acuities were negatively correlated in a large sample of subjects. Here, however, we observed a weak positive correlation between central (VisAcuity) and peripheral (CrowdSize) visual acuities ($r = 0.159$, $p = 0.129$; Table 2). It is therefore not completely unlikely that central and peripheral vision engage independently while playing video games, as was shown for reaching (Prado, Clavagnier, Otzenberger, Scheiber, Kennedy, & Perenin, 2005).

Other associations between visual paradigms and the players’ rank

Players with higher ranks were associated with stronger susceptibility to the Zöllner illusion (standardized path coefficient: −0.303; coefficient from the elastic net model: −0.106). Further investigation may closely examine this association.

In addition, we observed other associations. However, these explained only a small proportion of the variance of the ranks and were not always consistent across analyses (path model vs. elastic net model), suggesting that they may be unreliable. For example, faster reaction times when making saccades or searching for a target were associated with higher ranks (coefficients from the elastic net model: proSac: 0.080, VisSrch4: 0.011). Similarly, Bosten and colleagues (2017) reported that the time spent playing computer games significantly correlated with a factor for oculomotor speed.

Previous studies suggested that action video games are associated with better performance in certain perceptual tasks. Some of these associations could however not be replicated here. For example, VGPs were reported to perform significantly better than NVGPs at discriminating contracting, but not expanding, elements in a radial random dot kinematograms paradigm (Hutchinson & Stocks, 2013). The authors suggested

Downloaded from jov.arvojournals.org on 10/04/2021

that VGPs are more sensitive than NVGPs to visual characteristics, which are enhanced in gaming (e.g., contracting patterns) relative to those encountered in the real world (e.g., expanding patterns). Here, we did not observe any significant association between the players' rank and the performance in a radial random dot kinematograms (RDK rad). Note, however, that both contracting and expanding conditions were considered together in the RDK rad variable. In contrast, we observed that the performance in the horizontal random dot kinematograms (RDK hor) significantly loaded on the players' rank (standardized path coefficient: −0.212), even though the correlation between the two was only weak and nonsignificant ($r = 0.040$ $p = 0.706$; Table 2).

Li and colleagues (2009) reported that contrast sensitivity at intermediate and higher spatial frequencies was enhanced after action video game training, which suggests that high-rank players may have better sensitivity to contrast than low-rank players. However, we did not observe any significant association between contrast sensitivity and the players' rank. It seems unlikely that the spatial frequency used here (4.0 cy/arcdeg) was too low to find an effect, since a small but significant effect was previously reported with a spatial frequency of 3.0 cycles per degree. Likewise, while Li and colleagues (2010) previously reported that playing action game reduces the effects of backward masking, no such association was observed in the present investigation.

Similarly, we did not observe any significant association between the players' rank and perceptual speed (ReacTime), contrary to Dye and colleagues (2009). Importantly, we tested only gamers ranging from beginners to experts but not non-gamers, contrary to most previous studies, which may explain the discrepancies in the results. For example, it may be that training video game improves contrast detection and perceptual speed in NVGs but does not further improve with additional training.

Gaming variables

Among the six gaming variables that were extracted, four were retained in the elastic net model (Shoot, Spray, Track, and Flick) and two showed significant standardized path coefficients (Shoot and Track). To estimate to which extent the players' rank can be predicted from the performance in the six gaming variables, we computed another multiple regression model, in which only the gaming variables loaded on the actual CS:GO ranks. The six gaming variables accounted for 48% of the variance of the actual CS:GO ranks. Note that extracting more gaming variables could have resulted in a larger proportion of the variance explained. However, our results highlight that the Shoot and Track variables (which showed up in both path and elastic net models) are building blocks for the game.

Questionnaire variables

De Hesselle, Rozgonjuk, Sindermann, Pontes, and Montag (2021) reported that lower conscientiousness, extraversion, and agreeableness were significantly associated with more time spent gaming. Here, the associations between personality traits and gaming variables or the players' rank were in general weak. However, two personality traits showed significant associations with the players' rank in the path and elastic net models, namely the HEXACO CO and HEXACO OP. Results suggest that players with low scores in conscientiousness (HEXACO CO) and openness to experience (HEXACO OP) tend to have higher ranks.

Limitations

While we only tested CS:GO players, our results may hold true for other action video games. Importantly, the present investigation does not allow us to claim that high-rank CS:GO players develop specific visual skills while playing, such as better visual acuity in the periphery. Neither can we infer from the data that specific visual skills or personality traits are required to become an excellent player. However, a reliable dose-response effect in intervention studies was suggested as evidence for a causal effect of action video gaming on perception (Chopin et al., 2019). Although we are not able to infer causality, our experimental design avoids the methodological shortcomings inherent to intervention studies (Boot et al., 2011), such as differential placebo effects driven by the treatment versus control interventions (e.g., Tetris-trained participants may predict that they will have a better post-training performance in a mental rotation task). Likewise, all participants were active gamers, reducing the risks of strategy changes impacting our results. We cannot exclude gender-specific effects since only male participants took part in the present study (e.g., gender disparity in mental rotation ability decreased following video game training; see Feng et al., 2007).

Not all paradigms that we classified as visual are purely visual. Indeed, some paradigms also tap into more cognitive aspects, such as inhibition and attention. Hence, it may be that the significant associations with the players' rank are related to a complex interaction between visual perception and cognition.

Importantly, power may be an issue given the large number of variables extracted and the moderate sample size. Hence, results must be considered with caution and replicated. False-positive results (i.e., spurious

Downloaded from jov.arvojournals.org on 10/04/2021

associations) cannot be excluded given the large number of tests we computed. We do not have enough power to make conclusions on specific between-variable correlations in this study (Table 2). Instead, we aimed at showing that visual variables only poorly relate to each other in general. It is the pattern as a whole, which is important, not single, specific correlations.

Also, we do not have a measure of reliability for all variables, as not all variables were tested twice. Between-variable correlations may for example be underestimated because of poor or moderate reliability (Ackerman & Hambrick, 2020). However, note that reliabilities (as measured with intraclass correlations) were large for the central visual acuity (VisAcuity) and visual backward masking (VBM) variables, which suggests that variables measured with a staircase procedure show good reliability. Last, we considered the rank as a continuous variable, even though it is ordinal. As the distance between two ranks may not be constant, we consider this as a limitation of the present investigation.

Conclusions

To summarize, our results suggest that there is no strong common factor for visual perception in CS:GO players. However, the performance in some visual paradigms strongly predicts the players' rank. In particular, high-rank players seem to have better visual perception in the periphery, as measured with a crowding paradigm and the Honeycomb illusion, compared to lower-rank players. Even though causative relationships cannot be derived from these results, the present investigation gives clues about visual paradigms, which may be part of future training programs for esports.

Keywords: action video games, vision, common factor, prediction, visual acuity, honeycomb illusion

Acknowledgments

The authors thank Marc Repnow for technical support and Ophélie Favrod for helping with the selection of the visual paradigms.

Supported by the project "Basics of visual processing: from elements to figures" (project no. 320030_176153/1) of the Swiss National Science Foundation (SNSF) and by a National Centre of Competence in Research (NCCR Synapsy) grant from the SNSF (51NF40-185897).

Commercial relationships: AB, AM, and CC were employed at the company Logitech Europe S.A., a manufacturer of computer peripherals and software, in the R&D department.

Corresponding author: Aline F. Cretenoud.
Email: aline.cretenoud@gmail.com.
Address: Ecole Polytechnique Fédérale de Lausanne (EPFL), Laboratory of, Psychophysics, Brain Mind Institute, 1015 Lausanne, Switzerland.

References

Aberg, K. C., & Herzog, M. H. (2009). Interleaving bisection stimuli - randomly or in sequence - does not disrupt perceptual learning, it just makes it more difficult. *Vision Research, 49*(21), 2591–2598, https://doi.org/10.1016/j.visres.2009.07.006.

Ackerman, P. L., & Hambrick, D. Z. (2020). A primer on assessing intelligence in laboratory studies. *Intelligence, 80*, 101440, https://doi.org/10.1016/j.intell.2020.101440.

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*, 340–345.

Bach, M. (1996). The Freiburg Visual Acuity Test-Automatic Measurement of Visual Acuity. *Optometry and Vision Science, 73*(1), 49–53.

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science, 218*(4573), 697–698, https://doi.org/10.1126/science.7134968.

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research, 27*(6), 953–965, https://doi.org/10.1016/0042-6989(87)90011-3.

Bargary, G., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Hogg, R. E., & Mollon, J. D. (2017). Individual differences in human eye movements: An oculomotor signature? *Vision Research, 141*, 157–169, https://doi.org/10.1016/j.visres.2017.03.001.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17, https://doi.org/10.1023/A:1005653411471.

Bavelier, D., Shawn Green, C., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: Learning to learn and action video games. *Annual Review of Neuroscience, 35*, 391–416, https://doi.org/10.1146/annurev-neuro-060909-152832.

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Shawn Green, C., & Bavelier, D. (2018).

Downloaded from jov.arvojournals.org on 10/04/2021

Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77–110, https://doi.org/10.1037/bul0000168.

Bertamini, M., Cretenoud, A. F., & Herzog, M. H. (2019). Exploring the Extent in the Visual Field of the Honeycomb and Extinction Illusions. *I-Perception, 10*(4), 1–19, https://doi.org/10.1177/2041669519854784.

Bertamini, M., Herzog, M. H., & Bruno, N. (2016). The Honeycomb illusion: Uniform textures not perceived as such. *I-Perception, 7*(4), 1–15, https://doi.org/10.1177/2041669516660727.

Boot, W. R., Blakely, D. P., & Simons, D. J. (2011). Do Action Video Games Improve Perception and Cognition? *Frontiers in Psychology, 2*(226), 1–6, https://doi.org/10.3389/fpsyg.2011.00226.

Bosten, J. M., Goodbourn, P. T., Bargary, G., Verhallen, R., Lawrance-Owen, A. J., Hogg, R. E., … Mollon, J. D. (2017). An exploratory factor analysis of visual performance in a large population. *Vision Research, 141*, 303–316, https://doi.org/10.1016/j.visres.2017.02.005.

Bosten, J. M., & Mollon, J. D. (2010). Is there a general trait of susceptibility to simultaneous contrast? *Vision Research, 50*(17), 1656–1664, https://doi.org/10.1016/j.visres.2010.05.012.

Campbell, M. J., Toth, A. J., Moran, A. P., Kowal, M., & Exton, C. (2018). eSports: A new window on neurocognitive expertise? *Progress in Brain Research, 240*, 161–174, https://doi.org/10.1016/bs.pbr.2018.09.006.

Cappe, C., Clarke, A., Mohr, C., & Herzog, M. H. (2014). Is there a common factor for vision? *Journal of Vision, 14*(8), 1–11, https://doi.org/10.1167/14.8.4.

Chopin, A., Bediou, B., & Bavelier, D. (2019). Altering perception: the case of action video gaming. *Current Opinion in Psychology, 29*, 168–173, https://doi.org/10.1016/j.copsyc.2019.03.004.

Chung, S. T. L. (2013). Cortical reorganization after long-term adaptation to retinal lesions in humans. *Journal of Neuroscience, 33*(46), 18080–18086, https://doi.org/10.1523/JNEUROSCI.2764-13.2013.

Clark, K., Fleck, M. S., & Mitroff, S. R. (2011). Enhanced change detection performance reveals improved strategy use in avid action video game players. *Acta Psychologica, 136*(1), 67–72, https://doi.org/10.1016/j.actpsy.2010.10.003.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis : Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Education, 10*(7), 86–99, https://doi.org/10.1.1.110.9154.

Cretenoud, A. F., Francis, G., & Herzog, M. H. (2020). When illusions merge. *Journal of Vision, 20*(8), 1–15, https://doi.org/10.1167/JOV.20.8.12.

Cretenoud, A. F., Grzeczkowski, L., Bertamini, M., & Herzog, M. H. (2020). Individual differences in the Müller-Lyer and Ponzo illusions are stable across different contexts. *Journal of Vision, 20*(6), 1–14, https://doi.org/10.1167/JOV.20.6.4.

Cretenoud, A. F., Karimpur, H., Grzeczkowski, L., Francis, G., Hamburger, K., & Herzog, M. H. (2019). Factors underlying visual illusions are illusion-specific but not feature-specific. *Journal of Vision, 19*(14), 1–21, https://doi.org/10.1167/19.14.12.

da Cruz, J. R., Shaqiri, A., Roinishvili, M., Favrod, O., Chkonia, E., Brand, A., Figueiredo, P., … Herzog, M. H. (2020). Neural Compensation Mechanisms of Siblings of Schizophrenia Patients as Revealed by High-Density EEG. *Schizophrenia Bulletin, 46*(4), 1009–1018, https://doi.org/10.1093/schbul/sbz133.

de Hesselle, L. C., Rozgonjuk, D., Sindermann, C., Pontes, H. M., & Montag, C. (2021). The associations between Big Five personality traits, gaming motives, and self-reported time spent gaming. *Personality and Individual Differences, 171*, 110483, https://doi.org/10.1016/j.paid.2020.110483.

Dye, M. W. G., Green, C. S., & Bavelier, D. (2009). Increasing Speed of Processing With Action Video Games. *Current Directions in Psychological Science, 18*(6), 321–326, https://doi.org/10.1111/j.1467-8721.2009.01660.x.

Edwards, M., & Badcock, D. R. (1995). Global Motion Perception: No Interaction Between the First-and Second-order Motion Pathways. *Vision Research, 35*(18), 2589–2602.

Emery, K. J., Volbrecht, V. J., Peterzell, D. H., & Webster, M. A. (2017). Variations in normal color vision. VI. Factors underlying individual differences in hue scaling and their implications for models of color appearance. *Vision Research, 141*, 51–65, https://doi.org/10.1016/j.visres.2016.12.006.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology, 15*(2), 154–160, https://doi.org/10.1016/j.conb.2005.03.010.

Fahle, M., & Morgan, M. (1996). No transfer of perceptual learning between similar stimuli in the same retinal position. *Current Biology, 6*(3), 292–297, https://doi.org/10.1016/S0960-9822(02)00479-7.

Downloaded from jov.arvojournals.org on 10/04/2021

Faubert, J., & Sidebottom, L. (2012). Perceptual-cognitive training of athletes. *Journal of Clinical Sport Psychology, 6*(1), 85–102, https://doi.org/10.1123/jcsp.6.1.85.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science, 18*(10), 850–855, https://doi.org/10.1111/j.1467-9280.2007.01990.x.

Ferguson, C. J. (2007). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly, 78*(4), 309–316, https://doi.org/10.1007/s11126-007-9056-9.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78, https://doi.org/10.1016/j.paid.2016.06.069.

Green, C. S., & Bavelier, D. (2006). Enumeration versus multiple object tracking: the case of action video game players. *Cognition, 101*(1), 217–245, https://doi.org/10.1016/j.cognition.2005.10.004.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88–94, https://doi.org/10.1111/j.1467-9280.2007.01853.x.

Greenwood, J. A., Szinte, M., Sayim, B., & Cavanagh, P. (2017). Variations in crowding, saccadic precision, and spatial localization reveal the shared topology of spatial vision. *Proceedings of the National Academy of Sciences of the United States of America, 114*(17), E3573–E3582, https://doi.org/10.1073/pnas.1615504114.

Grzeczkowski, L., Clarke, A. M., Francis, G., Mast, F. W., & Herzog, M. H. (2017). About individual differences in vision. *Vision Research, 141*, 282–292, https://doi.org/10.1016/j.visres.2016.10.006.

Grzeczkowski, L., Cretenoud, A. F., Herzog, M. H., & Mast, F. W. (2017). Perceptual learning is specific beyond vision and decision making. *Journal of Vision, 17*(6), 1–11, https://doi.org/10.1167/17.6.6.

Grzeczkowski, L., Cretenoud, A. F., Mast, F. W., & Herzog, M. H. (2019). Motor response specificity in perceptual learning and its release by double training. *Journal of Vision, 19*(6), 1–14, https://doi.org/10.1167/19.6.4.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2018). *Multivariate data analysis* (8th ed.). Andover, UK: Cengage Learning EMEA.

Harris, P. B., & Houston, J. M. (2010). A Reliability Analysis of the Revised Competitiveness Index. *Psychological Reports, 106*(3), 870–874, https://doi.org/10.2466/pr0.106.3.870-874.

Herzog, M. H., & Koch, C. (2001). Seeing properties of an invisible object: Feature inheritance and shine-through. *Proceedings of the National Academy of Sciences of the United States of America, 98*(7), 4271–4275, https://doi.org/10.1073/pnas.071047498.

Herzog, M. H., Kopmann, S., & Brand, A. (2004). Intact figure-ground segmentation in schizophrenia. *Psychiatry Research, 129*(1), 55–63, https://doi.org/10.1016/j.psychres.2004.06.008.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4*(1), 11–26, https://doi.org/10.1080/17470215208416600.

Hutchinson, C. V, & Stocks, R. (2013). Selectively Enhanced Motion Perception in Core Video Gamers. *Perception, 42*(6), 675–677, https://doi.org/10.1068/p7411.

Iglewicz, B., & Hoaglin, D. (1993). Volume 16: how to detect and handle outliers. In *The ASQC basic references in quality control: statistical techniques*. Milwaukee: Asq Press.

Jackson, D. N. (1974). *Personality Research Form Manual*. London, Ontario: Research Psychologists Press.

Kaiser, H. F. (1970). A Second-Generation Little Jiffy. *Psychometrika, 35*, 401–415.

Knudson, D., & Kluka, D. A. (1997). The Impact of Vision and Vision Training on Sport Performance. *Journal of Physical Education, 68*(4), 17–24, https://doi.org/10.1080/07303084.1997.10604922.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.

Kowal, M., Toth, A. J., Exton, C., & Campbell, M. J. (2018). Different cognitive abilities displayed by action video gamers and non-gamers. *Computers in Human Behavior, 88*, 255–262, https://doi.org/10.1016/j.chb.2018.07.010.

Lahav, K., Levkovitch-Verbin, H., Belkin, M., Glovinsky, Y., & Polat, U. (2011). Reduced mesopic and photopic foveal contrast sensitivity in glaucoma. *Archives of Ophthalmology, 129*(1), 16–22, https://doi.org/10.1001/archophthalmol.2010.332.

Li, R., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience, 12*(5), 549–551, https://doi.org/10.1038/nn.2296.

Li, R., Polat, U., Scalzo, F., & Bavelier, D. (2010). Reducing backward masking through action game training. *Journal of Vision, 10*(14), 1–13, https://doi.org/10.1167/10.14.33.

Downloaded from jov.arvojournals.org on 10/04/2021

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate Practice and Performance in Music, Games, Sports, Education, and Professions: A Meta-Analysis. *Psychological Science, 25*(8), 1608–1618, https://doi.org/10.1177/0956797614535810.

Mason, O., Linney, Y., & Claridge, G. (2005). Short scales for measuring schizotypy. *Schizophrenia Research, 78*(2–3), 293–296, https://doi.org/10.1016/j.schres.2005.06.020.

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research, 141*, 4–15, https://doi.org/10.1016/j.visres.2017.11.001.

Nazhif Rizani, M., & Iida, H. (2018). Analysis of Counter-Strike: Global Offensive. *Proceedings of 2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 373–378, https://doi.org/10.1109/ICECOS.2018.8605213

Newsome, W. T., & Park, E. B. (1988). A Selective Impairment of Motion Perception Following Lesions of the Middle Temporal Visual Area (MT). *The Journal of Neuroscience, 8*(6), 2201–2211.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

Patino, C. M., McKean-Cowdin, R., Azen, S. P., Allison, J. C., Choudhury, F., & Varma, R. (2010). Central and Peripheral Visual Impairment and the Risk of Falls and Falls with Injury. *Ophthalmology, 117*(2), 206, https://doi.org/10.1016/j.ophtha.2009.06.063.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Peterzell, D. H. (2016). Discovering Sensory Processes Using Individual Differences : A Review and Factor Analytic Manifesto. *Electronic Imaging, 2016*(16), 1–11, https://doi.org/10.2352/ISSN.2470-1173.2016.16HVEI-112.

Peterzell, D. H., Schefrin, B. E., Tregear, S. J., & Werner, J. S. (2000). Spatial frequency tuned covariance channels underlying scotopic contrast sensitivity. In *Vision Science and its Applications: OSA Technical Digest*. Washington, DC: Optical Society of America.

Prado, J., Clavagnier, S., Otzenberger, H., Scheiber, C., Kennedy, H., & Perenin, M. T. (2005). Two cortical systems for reaching in central and peripheral vision. *Neuron, 48*(5), 849–858, https://doi.org/10.1016/j.neuron.2005.10.010.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, https://www.r-project.org/.

Roinishvili, M., Chkonia, E., Stroux, A., Brand, A., & Herzog, M. H. (2011). Combining vernier acuity and visual backward masking as a sensitive test for visual temporal deficits in aging research. *Vision Research, 51*(4), 417–423, https://doi.org/10.1016/j.visres.2010.12.011.

Schoups, A. A., Vogels, R., & Orban, G. A. (1995). Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity. *The Journal of Physiology, 483*(3), 797–810, https://doi.org/10.1113/jphysiol.1995.sp020623.

Shawn Green, C., Li, R., & Bavelier, D. (2010). Perceptual Learning During Action Video Game Playing. *Topics in Cognitive Science, 2*(2), 202–216, https://doi.org/10.1111/j.1756-8765.2009.01054.x.

Shawn Green, C., Sugarman, M. A., Medford, K., Klobusicky, E., & Bavelier, D. (2012). The effect of action video game experience on task-switching. *Computers in Human Behavior, 28*(3), 984–994, https://doi.org/10.1016/j.chb.2011.12.020.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428, https://doi.org/10.1037/0033-2909.86.2.420.

Smither, R. D., & Houston, J. M. (1992). The Nature of Competitiveness: The Development and Validation of the Competitiveness Index. *Educational and Psychological Measurement, 52*(2), 407–418, https://doi.org/10.1177/0013164492052002016.

Sowden, P. T., Rose, D., & Davies, I. R. L. (2002). Perceptual learning of luminance contrast detection: Specific for spatial frequency and retinal location but not orientation. *Vision Research, 42*(10), 1249–1258, https://doi.org/10.1016/S0042-6989(02)00019-6.

Spang, K., Grimsen, C., Herzog, M. H., & Fahle, M. (2010). Orientation specificity of learning vernier discriminations. *Vision Research, 50*(4), 479–485, https://doi.org/10.1016/j.visres.2009.12.008.

Spence, I., & Feng, J. (2010). Video Games and Spatial Cognition. *Review of General Psychology, 14*(2), 92–104, https://doi.org/10.1037/a0019491.

Spinella, M. (2007). Normative data and a short form of the Barratt Impulsiveness Scale. *International Journal of Neuroscience, 117*(3), 359–368, https://doi.org/10.1080/00207450600588881.

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions.

Downloaded from jov.arvojournals.org on 10/04/2021

The Journal of the Acoustical Society of America, 41(4), 782–787, https://doi.org/https://doi.org/10.1121/1.1910407.

Tibber, M. S., Guedes, A., & Shepherd, A. J. (2006). Orientation Discrimination and Contrast Detection Thresholds in Migraine for Cardinal and Oblique Angles. *Investigative Opthalmology & Visual Science, 47*(12), 5599–5604, https://doi.org/10.1167/iovs.06-0640.

Tulver, K. (2019). The factorial structure of individual differences in visual perception. *Consciousness and Cognition, 73*, 1–8, https://doi.org/https://doi.org/10.1016/j.concog.2019.102762.

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research, 141*, 217–227, https://doi.org/10.1016/j.visres.2016.12.014.

Wagner, M. G. (2006). On the Scientific Relevance of eSports. In *International Conference on Internet Computing,* 437–442.

Watson, A. B., & Pelli, D. G. (1979). *The QUEST staircase procedure. Applied Vision Association Newsletter, 14*, 6–7.

Watson, A. B., & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113–120, https://doi.org/10.3758/BF03202828.

Xiao, L. Q., Zhang, J. Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete Transfer of Perceptual Learning across Retinal Locations Enabled by Double Training. *Current Biology, 18*(24), 1922–1926, https://doi.org/10.1016/j.cub.2008.10.030.

Yashar, A., Wu, X., Chen, J., & Carrasco, M. (2019). Crowding and Binding: Not All Feature Dimensions Behave in the Same Way. *Psychological Science, 30*(10), 1533–1546, https://doi.org/10.1177/0956797619870779.

Yu, C., Klein, S. A., & Levi, D. M. (2004). Perceptual learning in contrast discrimination and the (minimal) role of context. *Journal of Vision, 4*(3), 169–182, https://doi.org/10.1167/4.3.4.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320, https://doi.org/10.1111/J.1467-9868.2005.00503.X.

Downloaded from jov.arvojournals.org on 10/04/2021

Supplementary File – How do visual skills relate to action video game performance?

Aline F. Cretenoud, Arthur Barakat, Alain Milliet, Oh-Hyeon Choung, Marco Bertamini, Christophe Constantin, & Michael H. Herzog

1. Supplementary Figures

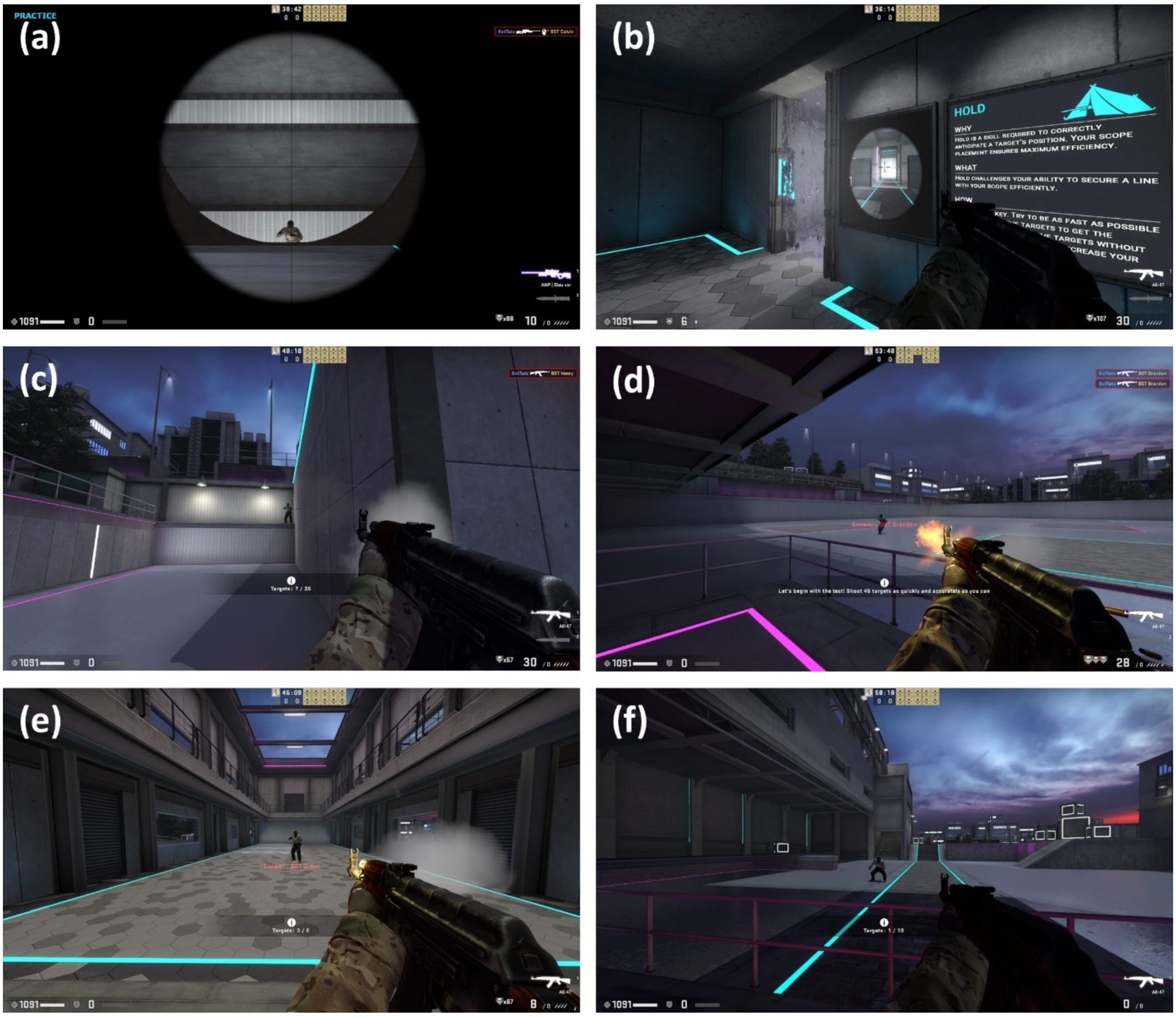


Figure S1. Screenshots of the (a) flick, (b) hold, (c) peek, (d) shoot, (e) spray, and (f) track CS:GO mini-games, which are publicly available on playmaster.gg.

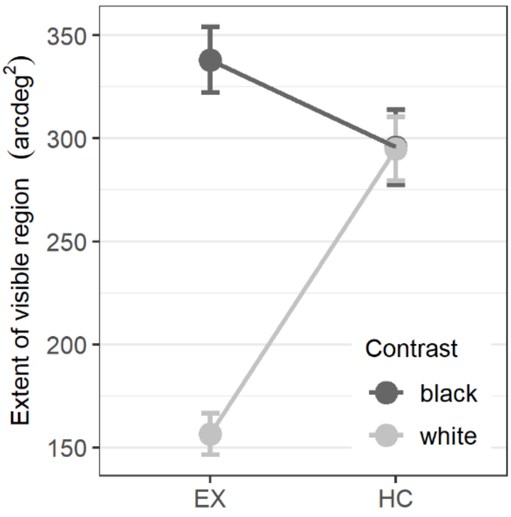


Figure S2. Mean extent of the region (in units of arcdeg^2) in which barbs (Honeycomb, HC) and disks (Extinction, EX) were perceptible as a function of the contrast polarity (black or white) in the HC/EX paradigm. Error bars show standard errors (*SE*).

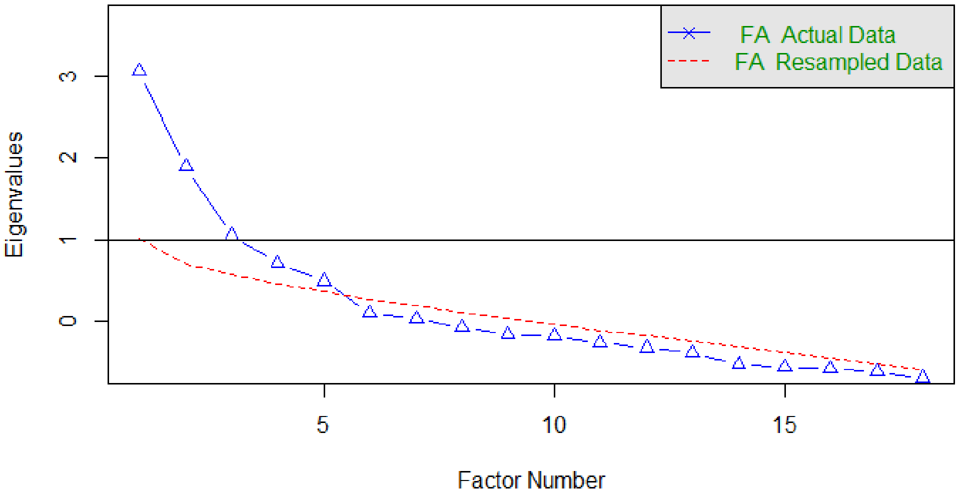


Figure S3. Scree plot from an exploratory factor analysis on the visual variables only. A common factor analysis was computed to extract the factors. Eigenvalues are shown for the actual data (in blue) and for resampled data (in red). A three-factor model is suggested by scree plot inspection, while a parallel analysis suggested a five-factor model.

2. Supplementary Tables

Table S1. Conversion from ordinal CS:GO ranks to numerical equivalents

Ordinal	Numerical
Silver I	1
Silver II	2
Silver III	3
Silver IV	4
Silver Elite	5
Silver Elite Master	6
Gold Nova I	7
Gold Nova II	8
Golda Nova III	9
Gold Nova Master	10
Master Guardian I	11
Master Guardian II	12
Master Guardian Elite	13
Distinguished Master Guardian	14
Legendary Eagle	15
Legendary Eagle Master	16
Supreme Master First Class	17
The Global Elite	18

Table S2. Reliability estimates expressed as an intraclass coefficient (ICC) of type (3,1) for each variable extracted from the HC/EX and Illusion paradigms, visual acuity (VisAcuity) and visual backward masking (VBM). Highlighted in orange are the variables, which showed a 95% confidence interval of the ICC including 0.5 (or below), indicating poor reliability according to Koo and Li (2016).

Paradigm	Variable	ICC		F test with true value 0			
		Coefficient	95% CI	F	df	p	pBonf
EX/HC	**EX black**	0.787	[0.713, 0.844]	8.398	[93, 93]	<0.001	<0.001
	EX white	0.848	[0.793, 0.890]	12.184	[93, 93]	<0.001	<0.001
	HC black	0.872	[0.824, 0.908]	14.647	[93, 93]	<0.001	<0.001
	HC white	0.822	[0.758, 0.870]	10.213	[93, 93]	<0.001	<0.001
Illusions	**CS left**	0.561	[0.432, 0.667]	3.551	[93, 93]	<0.001	<0.001
	CS right	0.569	[0.442, 0.674]	3.644	[93, 93]	<0.001	<0.001
	EB left	0.723	[0.630, 0.795]	6.213	[93, 93]	<0.001	<0.001
	EB right	0.605	[0.485, 0.703]	4.062	[93, 93]	<0.001	<0.001
	ML in	0.638	[0.525, 0.729]	4.522	[93, 93]	<0.001	<0.001
	ML out	0.528	[0.393, 0.640]	3.234	[93, 93]	<0.001	<0.001
	PD left	0.722	[0.629, 0.794]	6.196	[93, 93]	<0.001	<0.001
	PD right	0.666	[0.560, 0.751]	4.989	[93, 93]	<0.001	<0.001
	PZ down	0.639	[0.526, 0.729]	4.533	[93, 93]	<0.001	<0.001
	PZ up	0.371	[0.215, 0.509]	2.181	[93, 93]	<0.001	0.003
	TT left	0.317	[0.156, 0.462]	1.929	[93, 93]	<0.001	0.021
	TT right	0.399	[0.246, 0.533]	2.329	[93, 93]	<0.001	<0.001
	VH hor	0.530	[0.395, 0.642]	3.251	[93, 93]	<0.001	<0.001
	VH ver	0.756	[0.672, 0.820]	7.182	[93, 93]	<0.001	<0.001
	WH left	0.110	[-0.061, 0.275]	1.248	[93, 93]	0.144	1
	WH right	0.508	[0.370, 0.624]	3.062	[93, 93]	<0.001	<0.001
	ZN left	0.714	[0.620, 0.788]	6.001	[93, 93]	<0.001	<0.001
	ZN right	0.715	[0.620, 0.789]	6.009	[93, 93]	<0.001	<0.001
VisAcuity		0.811	[0.743, 0.862]	9.561	[93, 93]	<0.001	<0.001
VBM		0.847	[0.791, 0.889]	12.064	[93, 93]	<0.001	<0.001

Table S3. Statistics from Shapiro-Wilk tests and lambda (λ) exponent from an optimized Tukey transformation that maximizes normality for each visual (green), gaming (orange), and questionnaire (purple) variable according to the Shapiro-Wilk test

	CrowdSize	CrowdPeri	Contrast	HC black	HC white	EX black	EX white
SW statistic	0.925**	0.877**	0.732**	0.904**	0.947**	0.960*	0.820**
SW statistic Z	0.987	0.989	0.982	0.987	0.992	0.993	0.987
lambda (λ)	-0.70	0.45	0.10	0.40	0.40	0.55	0.45
	Poggendorff	**Zöllner**	**NBack**	**Orient**	**RDK hor**	**RDK rad**	**ReacTime**
SW statistic	0.986	0.977	0.960*	0.623**	0.765**	0.796**	0.923**
SW statistic Z	0.987	0.990	0.967*	0.988	0.971*	0.976	0.991
lambda (λ)	0.90	1.15	1.95	-0.35	0.05	0.05	-1.90
	proTravel	**proSac**	**antiTravel**	**antiSac**	**VisAcuity**	**VBM**	**VisSrch4**
SW statistic	0.554**	0.808**	0.486**	0.967*	0.982	0.820**	0.947**
SW statistic Z	0.989	0.985	0.983	0.993	0.986	0.968*	0.989
lambda (λ)	1.55	-0.75	0.70	-0.80	1.40	0.20	0.00
	VisSrch16	**Shoot**	**Spray**	**Track**	**Hold**	**Flick**	**Peek**
SW statistic	0.812**	0.975	0.983	0.955*	0.989	0.983	0.986
SW statistic Z	0.991	0.988	0.983	0.955*	0.991	0.990	0.987
lambda (λ)	-1.10	2.05	0.85	1.00	2.00	1.85	1.10
	AQ SS	**AQ AS**	**AQ AD**	**AQ CO**	**AQ IM**	**BIS NI**	**BIS MI**
SW statistic	0.943**	0.939**	0.956*	0.959*	0.950*	0.976	0.976
SW statistic Z	0.954*	0.940**	0.958*	0.959*	0.950*	0.977	0.980
lambda (λ)	0.80	1.10	1.20	1.00	1.05	0.75	0.60
	BIS AI	**CI EC**	**CI CO**	**HEXACO HH**	**HEXACO EM**	**HEXACO EX**	**HEXACO AG**
SW statistic	0.965*	0.932**	0.977	0.986	0.988	0.978	0.987
SW statistic Z	0.981	0.979	0.978	0.990	0.988	0.985	0.990
lambda (λ)	0.25	2.65	0.85	1.45	1.10	1.65	1.40
	HEXACO CO	**HEXACO OP**	**Handedness**	**O-LIFE UE**	**O-LIFE CD**	**O-LIFE IA**	**O-LIFE IN**
SW statistic	0.960*	0.975	0.725**	0.929**	0.958*	0.932**	0.950*
SW statistic Z	0.983	0.990	0.933**	0.954*	0.968*	0.945**	0.959*
lambda (λ)	2.35	1.95	3.55	0.75	0.75	0.80	0.80
	PRFd	**NbHours PerWeek**	**NbTotal Hours**	**Actual CS:GO rank**	**Best CS:GO rank**	**NbHours Sleep**	
SW statistic	0.974	0.828**	0.555**	0.929**	0.928**	0.946**	
SW statistic Z	0.976	0.975	0.973*	0.946**	0.930**	0.968*	
lambda (λ)	0.80	0.00	0.25	0.45	1.20	2.45	

The Shapiro-Wilk test was run both before (SW statistic) and after (SW statistic Z) data were transformed using a Tukey transformation with the optimized λ exponent (outliers were removed only in the visual variables). Significant statistics indicate a violation of the normality assumption. * $p < 0.05$, ** $p < 0.05/55$ (Bonferroni correction for multiple comparisons).

Table S4. Correlations between personality traits and all other variables expressed as correlation coefficients (Pearson's *r*). A color scale from blue to red shows effect sizes from *r* = -1 to *r* = 1. Numbers in italics indicate significant results without correction (α = 0.05). None of the significant correlations survived Bonferroni correction (α = 0.05/990; 990 correlations were computed in total, see Table 2 in the main text). Green, orange, and purple indicate visual, gaming, and questionnaire variables, respectively.

	AQ	**BIS**	**CI**	**HEXACO HH**	**HEXACO EM**	**HEXACO EX**	**HEXACO AG**	**HEXACO CO**	**HEXACO OP**	**Handedness**	**O-LIFE**	**PRFd**
CrowdSize	.18	-.09	-.02	.11	.18	-.10	-.15	.09	-.03	.01	.09	-.02
CrowdPeri	.03	.04	.03	.11	-.06	-.02	.05	.12	.16	-.02	-.08	.00
Contrast	-.10	.15	-.08	-.01	.18	.19	.19	-.12	.02	*-.26*	-.03	.01
HC black	*-.26*	-.09	*.25*	.02	*-.24*	.13	.00	*.23*	-.06	.19	-.07	.07
HC white	*-.21*	-.07	*.24*	.03	*-.27*	.11	.02	.19	-.05	*.23*	-.10	.02
EX black	-.14	-.19	*.23*	.12	*-.21*	.06	.06	.19	.01	*.20*	-.06	.01
EX white	-.16	-.04	*.29*	*.28*	*-.31*	.07	.02	.03	-.02	.13	-.08	-.03
Poggendorff	-.13	-.04	-.10	.12	-.02	.13	.06	.05	.13	-.15	-.08	.09
Zöllner	-.06	-.12	.13	-.13	-.06	-.01	.08	.11	.00	-.15	.07	.14
NBack	.03	.08	.08	.07	.00	.08	-.04	-.06	.08	-.07	.09	.07
Orient	-.14	.03	.06	*.22*	.04	-.12	.12	-.03	.10	-.01	.05	-.01
RDK hor	.04	-.04	-.19	-.05	.04	-.16	-.02	-.03	-.18	-.18	.06	-.09
RDK rad	-.13	-.02	-.09	.04	.05	-.08	.11	-.03	-.04	.05	-.07	.04
ReacTime	-.05	-.02	-.02	-.01	-.04	.00	-.10	.01	-.15	.06	-.06	.08
proTravel	-.05	.12	-.03	-.04	.02	-.03	.07	-.01	-.01	-.08	-.01	.03
proSac	-.18	-.02	-.17	-.02	-.03	.10	.05	-.12	-.11	.16	-.02	-.02
antiTravel	-.12	.19	.05	-.05	-.11	-.03	.03	-.06	.03	.01	.00	.06
antiSac	*-.25*	-.07	-.14	.05	-.03	.08	.04	.04	-.07	.14	*-.21*	.05
VisAcuity	-.08	-.09	-.16	.14	-.03	.02	.15	-.10	.03	*-.22*	-.09	-.05
VBM	-.04	.12	-.13	-.11	*.23*	.10	.18	*-.24*	-.06	*-.23*	.10	-.06
VisSrch4	*-.20*	.04	.08	-.15	.07	.14	-.14	-.02	-.05	-.08	.01	.03
VisSrch16	-.15	-.02	.00	-.11	.01	.04	-.14	.08	-.05	-.08	.09	.01
Shoot	-.06	.00	-.02	.18	-.11	.02	.02	.01	-.09	.04	-.06	-.15
Spray	.06	-.13	-.06	*.29*	-.19	-.17	.09	.11	-.13	.11	-.15	*-.25*
Track	-.06	.02	-.04	.15	-.12	-.06	.13	.02	-.14	-.06	-.18	-.13
Hold	-.12	-.14	.04	.19	-.01	.11	.03	-.07	.00	.00	*-.24*	.09
Flick	.07	.16	.01	.15	.03	-.08	.10	-.07	-.12	.10	.01	-.07
Peek	*-.30*	-.01	.07	.19	-.14	.16	*.24*	-.05	*.27*	.03	-.18	.09
NbHoursPerWeek	*-.23*	.13	.03	*.23*	-.09	.01	.07	-.05	-.08	.06	.00	*-.24*
NbTotalHours	-.13	.17	-.07	.16	-.17	.12	.08	-.07	*-.28*	.13	-.06	-.12
Actual CS:GO rank	-.03	.14	.00	*.23*	-.12	.04	.05	-.09	*-.28*	.13	-.05	*-.25*
Best CS:GO rank	-.07	.12	-.03	.18	-.10	.07	.05	-.07	*-.31*	.11	-.09	*-.21*
NbHoursSleep	-.02	.10	-.20	-.15	.08	-.14	.07	*-.23*	.00	.02	.03	-.12

Curriculum Vitae

PERSONAL INFORMATION

Name	Oh-Hyeon Choung
Date of Birth	7th February 1993
Citizenship	S. Korean
Address	EPFL SV BMI LPSY, Station 19 1015 Lausanne, Switzerland
Phone number	+41 78 721 57 91
Email	ohhyeon.choung@gmail.com

EDUCATION

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland 2017 –2021
Ph.D. in Neuroscience

Korea Advanced Institute of Science and Technology (KAIST), South Korea 2015 - 2017
M.S. in Neuroscience, College of Engineering (GPA: 3.9/4.3)

Korea Advanced Institute of Science and Technology (KAIST), South Korea 2011 - 2015
B.S. in Bio and Brain Engineering, College of Life Science and Bioengineering
GPA: 3.78/4.3 (Graduation with honor of *Cum Laude*)

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland 2014 - 2015
Exchange program, 1 semester

RESEARCH AND PERSONAL SKILLS

Programming skills: Python, R, MatLab, JavaScript, Java, C++, SQL

Deep learning & Machine Learning: Tensorflow, Keras, Pytorch, OpenCV, Scikit_learn, PySyft

Cloud computing: GCP, AWS

Human experiments: Psychtoolbox, Psychopy, Pavlovia, Gorilla, Prolific, Amazon Mturk

Medical imaging (fMRI) analysis: SPM, AFNI, FSL, Customized a toolbox for small animal fMRI

PUBLICATIONS

Oh-Hyeon Choung, Einat Rashal, Michael H. Herzog**,** (*in prep*). "Basic gestalt rules cannot explain Uncrowding".

Nadia Ruthemann*, **Oh-Hyeon Choung***, & Herzog, M. H., (*in prep*). **"**What crowds in crowding?**".**

Oh-Hyeon Choung, Alban Bornet, Adrien Doerig, Michael H. Herzog, (2021). "Dissecting (un)Crowding", *Journal of vision,* 21(10), 10.

Harshitha Machiraju*, **Oh-Hyeon Choung***, Pascal Frossard, Michael. H Herzog, (2021). "Bio-inspired Robustness: A Review". *arXiv preprint arXiv:2103.09265*.

Alban Bornet, **Oh Hyeon Choung**, Adrien Doerig, David Whitney, Michael Herzog, & Mauro Manassi, (*accepted*). “Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing”. *Jornal of vision,* 21

Aline F. Cretenoud, Arthur Barakat, Alain Milliet, **Oh-Hyeon Choung**, Marco Bertamini, Christophe Constantin, & Michael H. Herzog, (2021). “How do visual skills relate to action video game performance?”. *Journal of vision*, 21(7), 10.

Adrien Doerig, Alban Bornet, **Oh-Hyeon Choung**, and Michael H. Herzog, (2020). "Crowding reveals fundamental differences in local vs. global processing in humans and machines.". *Vision Research*, 167, 39-45.

Marc M. Lauffs*, **Oh-Hyeon Choung***, Haluk Öğmen, Michael H. Herzog, & Dirk Kerzel, (2019). "Reference-frames in vision: Contributions of attentional tracking to non-retinotopic perception in the Ternus-Pikler display". *Journal of vision,* 19(12), 7.

Marc M. Lauffs, **Oh-Hyeon Choung**, Haluk Öğmen, & Michael H. Herzog, (2018). "Unconscious retinotopic motion processing affects non-retinotopic motion perception.". *Consciousness and cognition*, 62, 135-147.

Oh-Hyeon Choung, Sang Wan Lee, & Yong Jeong, (2017). "Exploring Feature Dimensions to Learn a New Policy in an Uninformed Reinforcement Learning Task." *Scientific reports*, 7(1), 1-12.

Kwangsun Yoo, Sun Ju Chung, Ho Sung Kim, **Oh-hyeon Choung**, Young-Beom Lee, Mi-Jung Kim, Sooyeoun You, & Yong Jeong, (2015), “Neural substrates of motor and non-motor symptoms in Parkinson’s disease: a resting FMRI study.” *PloS one,* 10(4), e0125455.

CONFERENCE PROCEEDINGS

Oh-Hyeon Choung, Adrien Doerig, Alban Bornet, & Michael H. Herzog, “Recurrent Architectures are Needed for Human-like Global Processing”. 33rd Neural Information Processing Systems 2019 (NeurIPS 2019) workshop (SVHRM 2019), Vancouver, Canada, December 13th, 2019, Poster

Oh-hyeon Choung, Einat Rashal, & Michael H. Herzog, “Basic Gestalt laws cannot explain Uncrowding”, in the 42nd European Conference on Visual Perception 2019, Leuven, Belgium, August 26, 2019,Presentation

Oh-hyeon Choung, Marc M. Lauffs, Haluk Öğmen, Dirk Kerzel, & Michael H. Herzog, “Competing unconscious reference-frames shape conscious motion perception”, in Visual Science Society 2019, St. Pete Beach, Florida, USA, May 19, 2019, Poster

Oh-hyeon Choung, Marc M. Lauffs, Haluk Öğmen, Michael H. Herzog, “How unconscious retinotopic processing influences conscious non-retinotopic perception”, in Visual Science Society 2018, St. Pete Beach, Florida, USA, May 19, 2018, Poster

Hyunsu Lee*, Jinhee Yoon*, **Oh-Hyeon Choung***, Sunghong Park, & Yong Jeong, “Resting-state functional connectivity changes in photothrombotic ischemic stroke rat model”, in Society for neuroscience 2017, Washington DC, USA, Nov 15, 2017, Poster

Oh-Hyeon Choung, Yong Jeong, “Sequential integration of task-related dimensional components during multi-dimensional reinforcement learning task”, in Neuroscience 2016, San Diego, CA, USA, Nov 15, 2016, Poster

Oh-Hyeon Choung, Yong Jeong, “Incremental dimensional exploratory reasoning under multi-dimensional environment”, in 25th Annual Computational Neuroscience Meeting: CNS 2016, Jeju, South Korea, Jul 7, 2016, Poster

* denotes equal contributions.

TEACHING EXPERIENCES

Research supervisor for Bachelor and Master students (3 students)	2019 – 2021
TA for Statistics and Experimental design	2017 – 2020
TA for Neuroscience	2015 – 2020
TA for Molecular and Cellular Biology	2015 – 2016
TA for Bioengineering Laboratory I - Labview	2015 – 2016

HONORS

Pioneer program **Aug 2014 - Jan 2015**
Providing students an opportunity to conduct research on personal research project under guidance of a foreign researcher
($2,500, Department of Bio and Brain Engineering, KAIST)

LOTTE Scholarship **Sep 2012 - Feb 2015**
Supporting students with strong academic performance in the science and engineering in their pursuit of undergraduate studies in Korea, full tuition & fees
($2,500 per semester, LOTTE Foundation)