Laboratorio de Neurociencias

FACULTAD DE CIENCIAS
UDELAR fcien.edu.uy

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Contextual modulation and segmentation of naturalistic textures in peripheral vision

Lic. Daniel Herrera

Co-tutor: Ruben Coen-Cagli

Tutor: Leonel Gómez

Conoció con tanta seguridad el lugar en que se encontraba cada cosa, que ella misma se olvidaba a veces de que estaba ciega.

Cien años de soledad, Gabriel García Márquez

Agradecimientos

Muchas personas me ayudaron a transitar este camino, de muchas formas distintas. En primer lugar agradezco a mis padres, que me dieron las condiciones necesarias y su apoyo incondicional para que desarrolle lo mejor posible la carrera científica. Agradezco también a mis abuelos Chiqui y Álvaro, por el apoyo que me dieron y por su orgullo del camino que elegí, quienes hoy estarían doblemente orgullosos, y agradezco al resto de mi familia por lo mismo. Agradezco a mis muchos amigos que me acompañaron y apoyaron, y en particular a quienes siempre estuvieron dispuestos a escuchar los detalles de mis aventuras científicas a pesar de no ser científicos, Jairo, Mateo y Martín, y a mis amigos de la licenciatura que hicieron lo mismo, a pesar de ser científicos, Mauri, Martina, Santiago, Guille y Flo. Y un doble agradecimiento a Martín por prestarme tan amablemente su casa en La Paloma para transitar los momentos más intensos del camino. Agradezco a Felicia, también por su apoyo incondicional, por ayudarme a sacar la cabeza de la ciencia, y por todo lo que me aguantó.

Agradezco a los amigos del piso 4 de Facultad de Ciencias, donde siempre nos divertimos mucho y donde aprendí mucho más que ciencia: Tony, Diego, Matías, Emi, Yuyo, Pelo, Ana, Musto, Francesco, Adri y Nati. Le agradezco a mis amigos de la Albert Einstein College of Medicine de quienes aprendí y que hicieron más disfrutable mi estadía en la Gran Manzana: Dylan, Jonathan y Aida.

Agradezco a mis colegas del GUIAD-COVID, quienes me ayudaron a desarrollar la demencial idea de colaborar desde mis conocimientos, muchos de ellos adquiridos en esta tesis, con el manejo de la pandemia de COVID-19 en Uruguay mientras hacía el doctorado, en particular a Maine, Paola, Álvaro, Matías y Héctor.

Agradezco especialmente a Leonel, que me aceptó como estudiante, y con su propuesta del tema de investigación y con la total libertad y confianza que me otorgó, dió inicio a este proyecto, y a Ruben, quien también me aceptó como estudiante, y con su minuciosidad y atención al detalle me ayudó a darle fin. De ambos aprendí mucho y quedo muy agradecido.

Finalmente, agradezco a quienes fueron imprescindibles para el desarrollo de esta tesis: quienes participaron como sujetos en mis experimentos. Entre ellos: mis hermanos Andrés y Sebastián, mis amigos Juanjo, Cola, Piria, Rana, Nato, Erik, Jairo, Mateo, Mauricio, Pato, Teo, Yoda, Flo, Guille, Martina, Felicia, Andrés Rubio, Daniela, mis amigos del laboratorio Diego, Matías, Tony, Emi, Felipe, Alfonso y Bruno.

Table of Contents

1. PREFACE

1.1) PREFACE

It is difficult to communicate to people what it is to study vision. Seeing is so easy and natural to us, that understanding how it may give rise to a whole field of scientific research requires a stretch of the imagination. Nonetheless, in my teaching experience, I find that when the problem is presented from a different perspective, its complexity becomes more evident. After explaining the analogy between a digital image and the pattern of photoreceptor activation in the retina, asking "how could we make a robot see?", together with some reflection about the problem of vision, moves the task from the mundane and obvious, to the realm of the miraculous. This reflects the change of perspective from contemplating a complex well functioning system, to contemplating a blank slate where this monumentous system must be pieced together. Both perspectives are essential and complementary, and while the former is the most frequent in neuroscience, the latter can open new ways of understanding.

Of course, the value of thinking about how a visual system (or any other cognitive system) may be built has long been recognized in brain and mind research, even if due to its absence. This is one of the main topics in the philosophy of David Marr's 1982 book "Vision": A Computational Investigation Into the Human Representation and Processing of Visual Information, which marks an age in brain and mind research. Referring to the pioneering neuroscience research that had taken place in the decades prior to this book, he says in Chapter 1:

> *"As one reflected on these sort of issues in the early 1970s, it gradually became clear that something important was missing that was not present in either of the disciplines of neurophysiology or psychophysics. The key observation is that neurophysiology and psychophysics have as their business to describe the behavior of cells or of subjects, but not to explain such behavior. What are the visual areas of*

the cerebral cortex actually doing? What are the problems in doing it that need explaining, and at what level of description should such explanations be sought?

The best way of finding out the difficulties of doing something is to try to do it, so at this point I moved to the Artificial Intelligence Laboratory at MIT…"

Since David Marr's book, the computational analysis of brains and minds has had tremendous growth. Many things changed, such as the advent of ever faster and cheaper computers and computer graphics, a new explosion of AI research, with deep ripples on neuroscience and cognitive science, new technologies that allow to record thousands of neurons from behaving animals, and to finely control the activity of neural populations, fMRI techniques that allow to measure brain activity in humans non-invasively, among others. Large conferences, journals and research centers are dedicated to the interaction between brains, minds and machines and the development of theory for neuroscience and intelligence. Computational and systems neuroscience, where computational models are used to understand the behavior and function of neural systems is likely one of today's fastest growing and most exciting scientific disciplines. In this context, the current work is a very tiny contribution to this exciting field.

The subject of this work is the psychophysical and computational study of visual texture perception, and more specifically, contextual modulation and segmentation in peripheral vision. Although this may sound like a somewhat niche topic, it is a natural intersection point given by the progress of different research problems with long traditions. For example, texture perception and the use of texture-like representations by the visual system were among the most studied topics in the advent of computer graphics in vision research, in the pioneering investigations of Bela Julesz (Bela Julesz 1962), and it has since been a very active field. To give historical context, this highly influential work using mathematical tools to study texture representations happened at around the time when Hubel and Wiesel were carrying out their groundbreaking physiological work. Also, the analysis of the behaviorally relevant information contained in textures, and its use for moving animals, figures prominently in the landmark work of psychologist James Gibson, where the need to study vision from the perspective of ecological behavior is put forward (Gibson 1958). On the other hand, the segmentation of the visual world, or conversely the grouping of its elements, was recognized as a fundamental aspect of perception already by the Gestalt school early in the 20th century (Wagemans et al. 2012). Despite being widely studied since the Gestalt period, and still being recognized as a key process in visual perception, grouping remains a somewhat mysterious and problematic process (it is closely related to the widely discussed *binding problem*) (Treisman 1996; Roelfsema 2006). Finally, contextual modulation, the phenomenon in which the context of a stimulus modifies the neural or perceptual responses it generates, is also recognized as a fundamental process of vision. After the initial “simplified” view of the primary visual cortex (V1) as the cells responding consistently to a given visual pattern in their receptive field, it was then recognized that these responses are affected by the context surrounding this receptive field, opening a new much more complex perspective on the function of V1 (Gilbert et al. 1996). Besides the phenomenological description of these contextual interactions (in V1 and many other brain areas), they have been associated with many processes, such as efficient coding, grouping, segmentation, predictive coding, among others (Coen-Cagli, Kohn, and Schwartz 2015; Zhaoping 1998; Rao and Ballard 1999; Malik and Perona 1990; Graham, Sutter, and Venkatesan 1993). Also, in perception, contextual modulation in peripheral vision (which occupies most of our visual field) has been recognized as a major limiting factor of our visual capabilities, and has also given rise to vast numbers of studies trying to understand its nature, in order to understand in what ways peripheral vision is limited (Bouma 1970). Thus, these three topics, texture perception, segmentation and contextual modulation, have all been pillars of vision research through the 20th and 21st century, and each has been the locus of much interaction between psychophysics, physiology and computational models.

In recent years, the links between these different topics have risen rapidly, and our study of the effects of segmentation on contextual modulation of texture perception in peripheral vision is the natural response to many questions that have emerged from this interaction. Specifically, we studied a type of textures that are defined in a very influential mathematical model, the Portilla-Simoncelli texture model (Portilla and Simoncelli 2000), that is a continuation of the pioneering work of Julesz. Besides continuing this long tradition of texture perception work, these textures have also provided the substrate stimuli for a very prolific line of research into the physiology of the visual cortex. These stimuli have allowed to analyze, in a principled and formalized way, the emergence of selectivity to complex stimuli in mid-level visual areas V2 and V4 from the input they receive from the primary visual cortex V1 (Freeman et al. 2013). Moreover, this texture model has also provided the basis for one of the main models of peripheral vision processing: the summary-statistics encoding model (SS model) (Rosenholtz 2016). This model of peripheral vision is proposed to provide a unifying account of several disparate phenomena, among which is the sometimes intricate workings of contextual modulation in peripheral vision. But this model of peripheral encoding has been challenged based on its apparent failures to account for important segmentation and grouping phenomena (Doerig et al. 2019). Thus, Portilla-Simoncelli textures provide a remarkable tool where perception, physiology, and mathematical models of sensory processing converge naturally, and that can allow to study from these three levels the interaction between segmentation, texture perception and contextual modulation. Therefore, although the present work focuses mostly on psychophysics, our experiments also heavily draw inspiration from the related physiological and computational literature. We also put an emphasis in interpreting the results from the perspective of computational models which serve to bridge the gap between perception and physiology. Finally, recent advances in artificial intelligence have renewed the interest in this field in the role of textures (Geirhos et al. 2018; L. A. Gatys, Ecker, and Bethge 2016), and recurrent processes related to grouping (Sabour, Frosst, and Hinton 2017; LaLonde and Bagci 2018) in visual perception. We hope that our tiny contribution of analyzing human perception from the perspective of these influential models helps in continuing to raise bridges between theory and experiment.

2. INTRODUCTION

2.1) PHYSIOLOGY OF THE EARLY VISUAL SYSTEM

2.1.1) The retina and Lateral Geniculate Nucleus

The transduction of light into neural activity occurs at the retina. There, specialized cells called *photoreceptors* contain pigments that change their configuration when impacted by light. This change in configuration releases a cascade of intracellular signaling that results in a graded modulation of the membrane potential of the cell. This change in membrane potential in the cell then results in changes in the rate of neurotransmitter release. Photoreceptors form a 2D sheet covering the retina, and this way, the 2D activity pattern of photoreceptors signals to the downstream visual system the patterns of light that fall onto the retina.

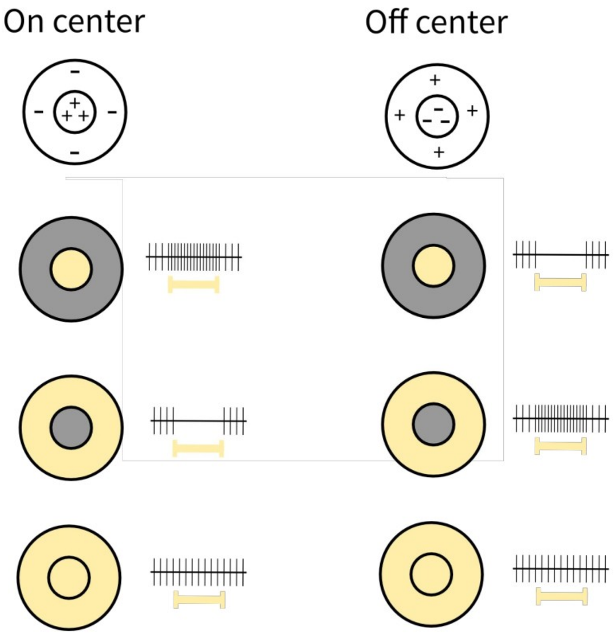


Figure 1. **Antagonistic center/surround receptive fields**. On the top row, a diagram of the excitatory and inhibitory structure of on-center (left column) and off-center (right column) receptive fields in the retina and lateral geniculate nucleus (LGN) is shown. On the next three rows, different light patterns used to stimulate the receptive field are shown. To the right of each stimulus, the responses of the two different neurons to these patterns are shown, with each vertical line indicating

an action potential of the neuron, and the horizontal yellow bar indicating the period during which the stimulus is shown. Image reproduced from *http://miladh.github.io/lgn-simulator/doc/recepfield.html*, GNU GPL v3.0 license.

Still in the retina, photoreceptors connect through synapses to a layer of bipolar cells, to which they communicate their outputs. This layer of bipolar cells sends their outputs to another type of neuron called *ganglion cells*, which are the output neurons of the retina. Because of the patterns of connectivity between these layers of neurons in the retina, ganglion cells are activated by specific patterns of light falling onto a given patch in the retina. The region of the visual field that activates a given cell, or more specifically the function that maps pattern of light to cell activity, can be called the *receptive field*[1] of the cell. Specifically, most retinal ganglion cells show antagonistic center/surround receptive fields, where either they are excited by light falling in the center of their receptive field and inhibited by light falling off the center, or vice versa **(Figure 1**). This receptive field configuration makes these neurons respond to changes in luminance in the visual input, rather than to homogeneous surfaces, since the latter stimulate similarly the excitatory and inhibitory components of the receptive field, thus canceling out. Furthermore, the size of the receptive fields changes between populations of ganglion cells, giving rise to channels processing the visual input at different scales. Like with photoreceptors, ganglion cells tile the retina, providing a 2D pattern of activation that signals the presence across the visual field of the patterns to which they are selective (Frisby and Stone 2010).

Then, ganglion cells send axons through the optic nerve to the Lateral Geniculate Nucleus (LGN). The LGN is a bilateral structure, and the LGN on each side receives projections from both retinas. Each LGN receives projections corresponding to the visual field from the opposite visual hemifield, that is, the left LGN processes the right visual hemifield, and vice versa (**Figure 2**). Although the LGN is a complex structure, its cells also show broadly a similar antagonistic center/surround structure in their receptive fields. Finally, the LGN cells project to the primary visual cortex (V1) in an ordered and retinotopic fashion, in which neighboring regions of V1 receive projections from neighboring areas of the visual field.

[1] The concept of receptive field is somewhat more complex than this and the term receptive field will be used somewhat more loosely than this given definition in the text.

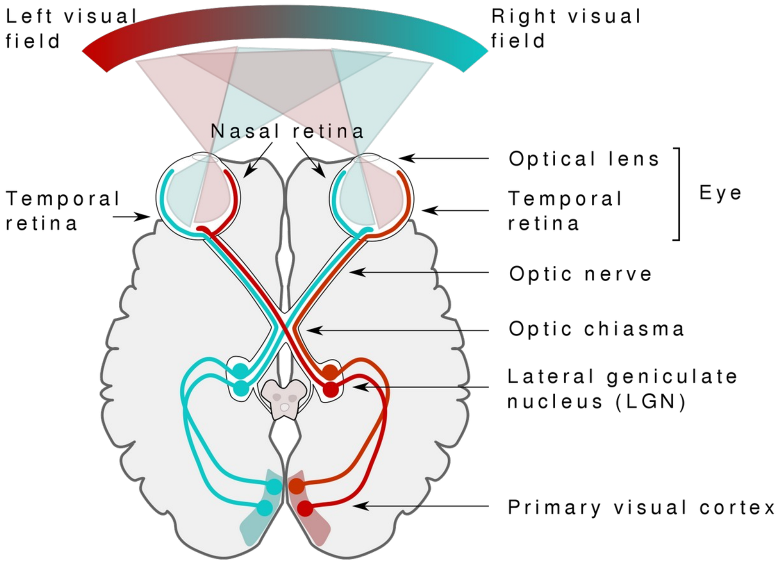


Figure 2. **Retina to cortex projections**. The visual information from the two visual hemifields is present in each retina. In the optic chiasma, the retinal projections corresponding to the contralateral visual hemifield of each eye continue to the LGN in the same side, while the retinal projections of the ipsilateral visual hemifield cross over to the LGN in the other hemisphere. This way, each LGN receives the projections from the contralateral visual hemifield. Then, each LGN sends the visual information to the ipsilateral primary visual cortex. Image by Miquel Perelló Nieto CC BY-SA 4.0.

2.1.2) Primary visual cortex: Receptive fields

The visual cortex V1 (also called striate cortex), located in the occipital lobe, is the first cortical processing stage of the visual input. Neural recordings of V1 cells show that the responses of these cells are tuned to many different dimensions of the visual input. One of the most studied and important tuning characteristics of V1 cells is their *orientation tuning*. As was shown originally by Hubel and Wiesel (Hubel and Wiesel 1959; 1968), V1 cells responses usually show specificity to oriented patterns of light in the retina, such as lines or edges. They respond maximally to a given orientation, with their responses declining as the stimulus is rotated further away from the optimal orientation (**Figure 3A**).

Hubel and Wiesel also distinguished between two types of cells in the visual cortex. The first are the so called *simple cells*. These have oriented receptive fields with excitatory and inhibitory regions, and their response is sensitive to the specific location of the oriented bar within their receptive fields. This cell may, for example, be excited as a bar of light with the proper orientation falls into the excitatory region of the receptive field, but inhibited when the bar falls into the inhibiting regions of the receptive field. The receptive fields of these cells are proposed to reflect a pattern of connections in which a given simple cell receives inputs from LGN cells whose receptive fields are aligned in the preferred orientation of the cell (**Figure 3B, top**) (Hubel and Wiesel 1962).

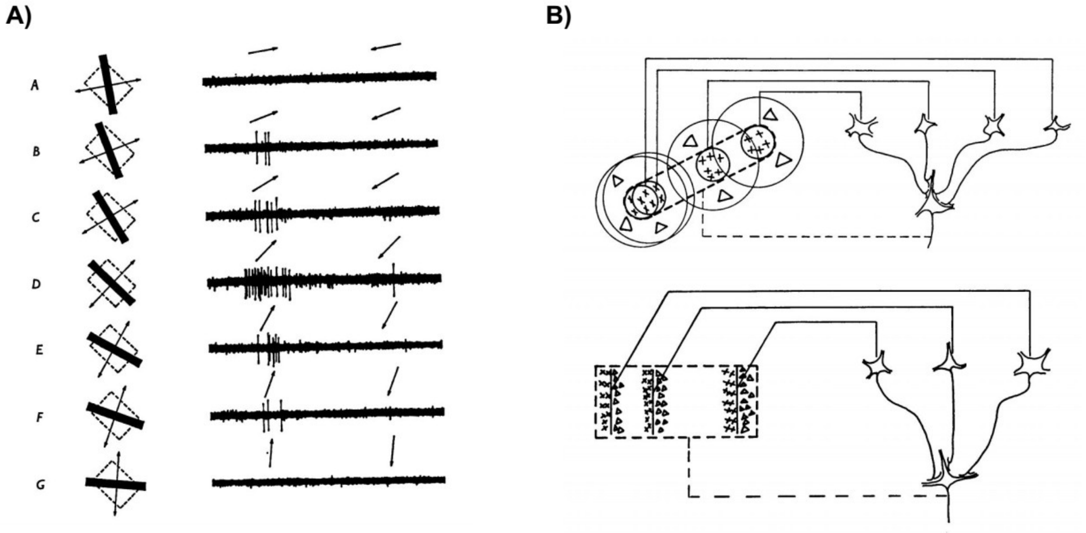


Figure 3. Orientation selectivity in V1. A) Recordings of a neuron from the primary visual cortex of a macaque monkey. To the left, the different stimuli displayed in the receptive field of the neuron are shown, and they constitute moving oriented bars. To the right, the traces of the recordings of the neuron show the spikes in response to each orientation and movement direction. Reproduced from (Hubel and Wiesel 1968). **B)** Diagram showing how a simple V1 cell may build its orientation selectivity by pooling together the outputs of several LGN neurons whose receptive fields are aligned in one direction (top), and how a complex V1 cell may build its location invariance by pooling several simple cell outputs that have the same orientation but varying positions and receptive field phases (bottom). Reproduced from (Hubel and Wiesel 1962)

The other major kind of cell Hubel and Wiesel recognized were the *complex cells*. Like simple cells, complex cells also show orientation tuning. But unlike simple cells, they do not show differentiated excitatory and inhibitory regions that can straightforwardly produce this orientation tuning. Rather, they are excited by oriented bars falling anywhere in their receptive field. They show thus position invariance (or also phase invariance). Complex cells achieve this by receiving inputs from multiple simple cells that are tuned to the same orientation, but that differ slightly the location or the polarity (i.e. phase) of their receptive field (**Figure 3B, bottom**) (Hubel and Wiesel 1962).

V1 cells also show other forms of tuning. Another important tuning dimension is the spatial frequency, or scale, of the oriented structure. While some cells respond to oriented structures of finer scale, or higher spatial frequency, others respond to coarser scales, or lower spatial frequencies (Mazer et al. 2002). This can be seen when using sinusoidal gratings in a display to stimulate recorded neurons, where for an optimal orientation of the grating, the response of the cell is maximal for a given spatial frequency, and decreases as the spatial frequency is changed. Lastly, V1 cells also show other types of tuning such as tuning to the eye of origin, direction of motion and color. V1 cells with specific receptive field properties tile the visual field, as has been described for the cells in the retina. Therefore, V1 also provides 2D activation maps that signal the presence of different features (i.e. different orientations, spatial frequencies, disparities, etc) across the visual scene.

2.1.3) Primary visual cortex: Contextual modulation and extraclassical receptive fields

Besides being driven by the light patterns that fall into their receptive fields, the response of V1 neurons is also modulated by the visual input falling outside their receptive fields, which by itself would not evoke responses in the neurons. This region surrounding the neurons receptive field, with the capacity to modulate its response, is called the extraclassical receptive field. We refer as *contextual modulation* to this phenomenon in which visual context modulates how neurons respond to the stimuli in their receptive fields[2].

As with the receptive field, the extraclassical receptive field is tuned to the characteristics of the visual input, with different surround patterns producing different modulations on a neurons response (Angelucci et al. 2017). There is a wide range of dimensions and details to which the contextual modulation of a neuron may be tuned, and these also vary between neurons. For example, a common observation

[2] Later, we will also define contextual modulation in similar terms for perception of the visual input. The term *contextual modulation* will therefore refer to both physiological and perceptual phenomena.

is that when stimuli in the surrounds are dissimilar to the stimulus in the receptive field (e.g. different orientations or spatial frequencies), they can facilitate the response of the neuron. Conversely, stimuli in the surround that are similar to the stimulus in the receptive field can suppress the response of a neuron. Nonetheless, in some cases surrounding stimuli similar to central stimuli have been reported to exert a facilitatory influence, for example, when they are all collinear (Angelucci et al. 2017). There is also interaction between different stimulus dimensions. For example, the sign of the contextual modulation effect, as well as the degree of orientation tuning, can depend on the contrast of the central and surrounding stimuli (Angelucci et al. 2017; C. A. Henry et al. 2013). Moreover, the sign of the contextual modulation as well as its tuning to the stimulus can change through time, reflecting different and interacting contextual modulation processes, associated with different neural mechanisms (Christopher A Henry et al. 2020). As can be appreciated from these examples, contextual modulation is quite intricate.

But besides the detailed descriptions obtained from studying contextual modulation with simple stimuli such as lines and gratings, contextual modulation in V1 neurons has also been studied using natural or naturalistic images. This venue of research is important for testing the tuning of contextual modulation to the natural structure of images, as well as to elucidate its possible functions in natural vision. For example, in a series of landmark studies registering the activity of V1 neurons in response to natural scenes, it was found that when the extraclassical receptive fields were stimulated with the context of the natural scenes, the responses of the neurons were much sparser (i.e. they responded to fewer images) (Vinje and Gallant 2000) and the neurons transmitted information more efficiently (Vinje and Gallant 2002). Other studies have shown that, when V1 neurons respond to natural images, the original natural image surrounds exert much stronger contextual modulation than surrounds that where distorted through phase-scrambling[3] to destroy their natural structure (Pecka et al. 2014; Guo et al. 2005; Coen-Cagli, Kohn, and Schwartz 2015). This has been argued to be due to the tuning or adaptation of contextual modulation to natural image structure.

[3] Phase-scrambling is a technique that maintains part of the natural structure of the image (up to second-order statistics) but discards the higher order structure. Phase-scrambled images lose their natural appearance but keep important properties of the original images. See **Section** 2.2.3 for further detail on phase scrambling.

Also, regarding the neural substrate for these contextual modulations, they emerge from three different sources (Christopher A Henry et al. 2020): from the feedforward input, from recurrent horizontal connections, and from feedback connections. The feedforward source arises, for example, by inheriting contextual modulation processes that occur in the retina and in the LGN. The horizontal connections contribution involves the connections between cells in V1, where cells are interconnected through short projections that also show tuning in their wiring patterns (Angelucci et al. 2017; 2002; Seriès, Lorenceau, and Frégnac 2003). The feedback surround modulation involves projections received from higher-level visual areas, which also show preference in their connectivity, and also exert contextual modulation effects across larger distances than horizontal connections (Angelucci et al. 2017; 2002; Poort et al. 2016).

Finally, it is worth mentioning a specific type of contextual modulation named divisive response normalization, in which the response of a neuron is normalized by the pooled responses of nearby neurons. This response normalization phenomenon has been used as one of the main explanatory models of contextual modulation phenomena in V1, and it has been proposed to be a canonical computation (Carandini and Heeger 2012). A canonical computation is a standard computation that is applied repeatedly across brain areas and neural systems, or a kind of elemental computation. In support of this hypothesis, divisive normalization has been observed across multiple neural systems (Carandini and Heeger 2012). As for many other neural computations, divisive normalization has been mostly studied on area V1, and thus research into contextual modulation in V1 serves as a guide for studying the functional principles of this computation in other neural systems.

2.1.4) Visual cortex V2

Visual cortex V2 is the area to which V1 sends the most projections, and it is the second largest visual area after V1 (Freeman et al. 2013; Sincich and Horton 2005). Following what has been discussed for the previous stages of the visual cascade, it could be expected that visual area V2 performs some straightforward combination of the different kinds of receptive fields in V1 (i.e. make different combinations of orientations and spatial frequencies to build a new kind of visual pattern) (Riesenhuber and Poggio 1999). But the question of what selectivity characterizes area V2 is still a developing one, with huge progress made in recent years.

V2 neurons have a strong orientation selectivity, as the earlier area V1, but processing characteristics have been proposed to emerge in V2. Some examples of these are: that V2 neurons respond to oriented edges defined by complex cues such as texture discontinuity, or illusory edges (Peterhans and Heydt 1993); that V2 cells have selectivity for more complex shapes, such as angles, and complex gratings such as polar gratings (Hegdé and Essen 2000); that orientation preference changes within the receptive field of some V2 cells, and that changes in orientation preference allow the encoding of combinations of orientations (Anzai, Peng, and Van Essen 2007), just to name a few.

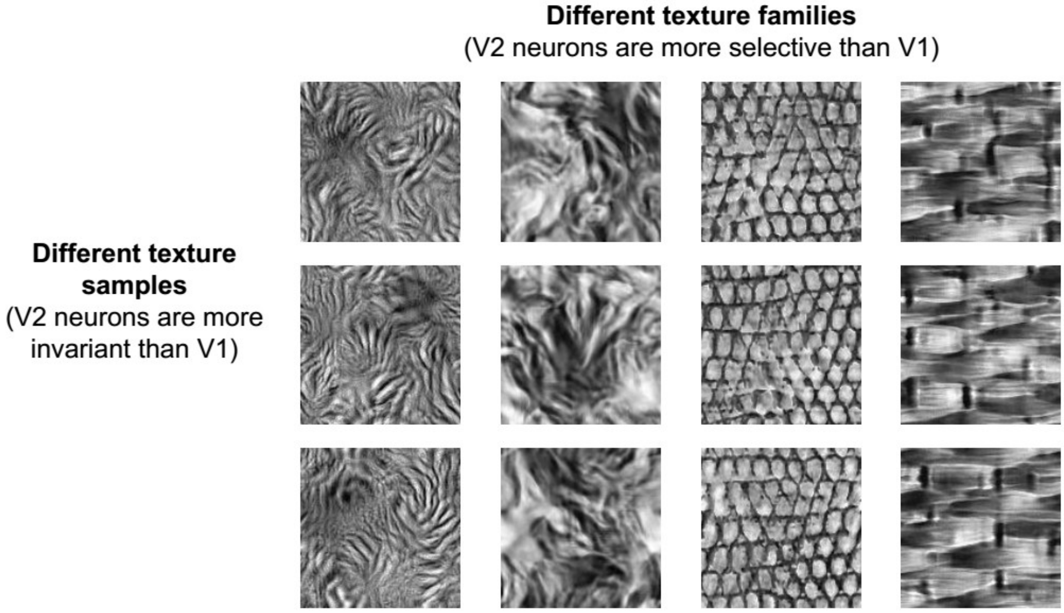

Different texture families
(V2 neurons are more selective than V1)
Different texture samples
(V2 neurons are more invariant than V1)

Figure 4. Local statistics selectivity in V2. Different textures synthesized with the PS model are shown in the image. Each column shows a different family of textures, each with its own set of values for the PS model statistics. Each row shows one different texture sample for the corresponding texture family, but all samples in a given column share the same set of statistics. That is, all the images in a given column have the same statistical distribution of V1-like features (i.e. the same correlations between V1-like filter activations across the image), but differ in the specific layout of these features. V2 neurons are more selective to the texture family and more invariant to the texture sample than V1 neurons.

But recently, another idea that has gained traction is that V2 neurons respond to the correlations between V1-like features within their receptive fields, or equivalently, that they respond to local texture properties (Freeman et al. 2013; Ziemba et al. 2016). In other words it is proposed that the activity of V2 neurons encodes the local texture in an image. This hypothesis, which followed from earlier perceptual work showing the potential relevance of summary-statistics representations in peripheral vision (Balas, Nakano, and Rosenholtz 2009; Jerome Y. Lettvin 1976), has been probed using *Portilla-Simoncelli (PS) textures* (Portilla and Simoncelli 2000), which are defined by a set of image statistics that are mostly characterized by the correlations between the outputs of V1-like filters (e.g. **Figure** 10, see **Section 2.2.2** further detail on PS textures).

One key result in this previous line of work is that V2 neurons show increased selectivity for PS statistics in texture stimuli compared to V1, while also showing increased invariance to the precise instantiation of the texture (Ziemba et al. 2016). Experiments testing this are based on the property of textures that a *family* of textures defined by a set of PS statistics can have various different instantiations or *samples* (**Figure 4**). While the textures samples in a given family have the same statistical relations between the V1 filter outputs, they may have very different distributions of these filter activations in space. For example, note in **Figure 4** that while the textures in a given column share the same appearance due to their shared statistical structure, they are completely different if compared pixel-by-pixel or patch by patch. Therefore, the observation that compared to V1, V2 neurons in macaque show increased selectivity for PS texture family while also showing increased invariance to the precise configuration of V1-like features shows the emergence of a texture-like representation in V2 (Ziemba et al. 2016). This is in line with previous physiological work (Rust and DiCarlo 2010) showing increased invariance and selectivity when going from V4 to area IT. It has also been shown that V2 neurons respond specifically to textures with naturalistic correlations between these V1-like features, since phase-scrambled patches of texture (which keep the same “amount” of V1-like features as the source texture but have no correlations between the V1-like features, see **Section 2.2.3**) drive V2 neurons more weakly than than PS textures synthesized with the statistical structure of natural images (Freeman et al. 2013).

2.1.5) V4 and the ventral stream:

After area V2, which is close to area V1 in the visual hierarchy, a major characteristic of the primate visual system starts to arise: the division between the temporal and the dorsal streams. This division is seen in the anatomical location of these regions, as well as in their patterns of connectivity, which show two clear clusters of areas beyond V2 (**Figure 5**) (Felleman DJ and Van Essen DC 1991; Markov et al. 2014). This anatomical division also reflects a functional division, with the ventral and dorsal streams being involved in different kinds of visual tasks. What is the actual nature of this division is somewhat still an open question, but there is a clear division between the kinds of tasks related to the two streams. Originally, it was proposed that the ventral stream is involved in visual recognition (e.g. recognizing faces, objects, scenes), while the dorsal stream is involved in spatial perception, including location and motion. This division resulted in the ventral and dorsal streams being labeled the ‘what’ and ‘where’ streams respectively (Goodale and Milner 1992). An alternative but somewhat related proposal is that the distinction is between the use of the outputs of the two systems: while the output of the ventral stream sustains perception of the environment, the dorsal stream guides actions in the environment (Goodale and Milner 1992).

Visual area V4 is another mid-level visual area. It forms part of the ventral pathway, and it receives projections from areas V1 and V2. What is relevant to us about area V4 is that the selectivity of V4 neurons to PS textures has also been studied in some detail. Following with the hierarchical processing scheme, by which each area responds to more complex patterns than the previous areas, it has been found that V4 neurons also show selectivity for PS statistics, and that this selectivity is stronger than for V2 neurons (Okazawa, Tajima, and Komatsu 2015; 2017). Although area V4 does much more than encoding local texture, as is reflected by the different kinds of stimuli modalities and configurations to which V4 shows selectivity (Roe et al. 2012; Pasupathy, Kim, and Popovkina 2019), this result shows that PS statistics are a powerful tool to probe the visual system.

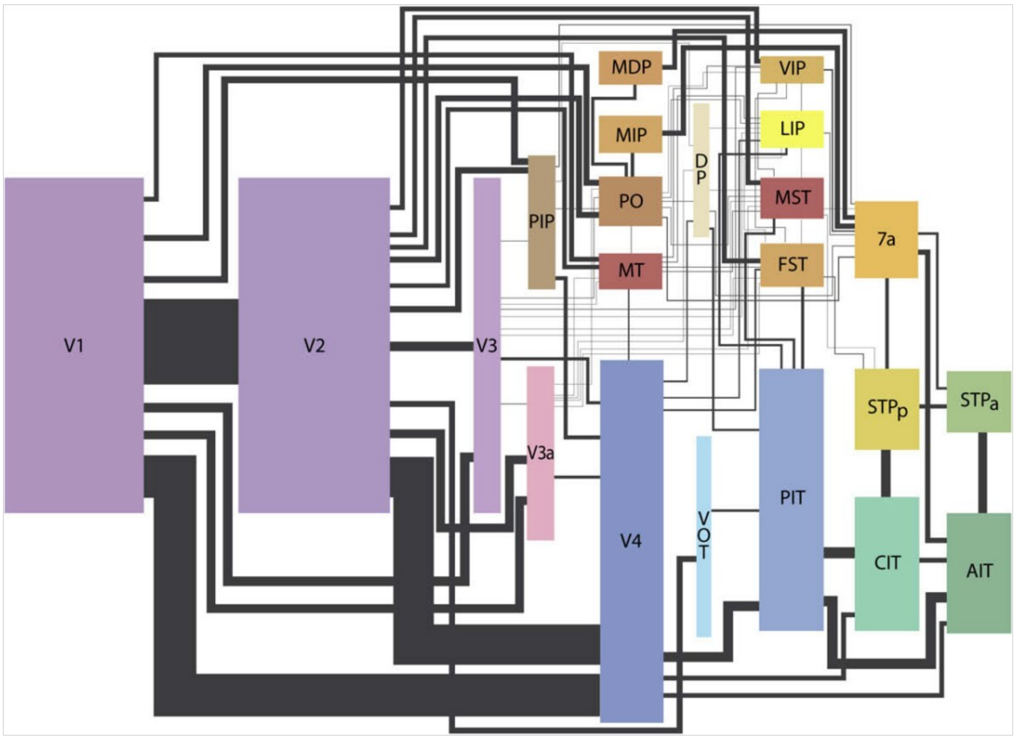


Figure 5. Visual system areas and connectivity. Map of the areas of the visual system, and their interconnectivity. Each rectangle represents a visual area, and each line connecting two rectangles represents a pathway connecting the two areas. The size of the rectangle of each visual area is proportional to its cortical area. The width of the line connecting each rectangle is proportional to the estimated connections between the two. We see how two clusters of areas are defined in this map, with the areas corresponding to the dorsal stream shown on the top of the diagram with shades of red and yellow, and the areas corresponding to the ventral stream shown on the bottom, with shades of blue and green. Reproduced from (Wallisch and Movshon 2008).

Downstream of area V4, several areas in the ventral stream have neurons with selectivity to complex objects, together with considerable invariance to factors such as contrast, viewpoint, size, or specific location in the visual field. In particular, the role of the inferior temporal (IT) cortex for object and pattern recognition was well established from lesion studies, but later, neurons that responded specifically to hands or faces with high invariance were found (Desimone et al. 1984). These neurons also have large receptive fields, in line with the idea that at each stage of the visual hierarchy, the integration of responses from earlier stages across visual space makes for larger receptive fields (Desimone et al. 1984). Despite individual neurons in these high-level areas not necessarily showing selectivity for object classes or semantic categories, the neuronal populations in these areas encode a high-level representation of image structure that allows these categories to be easily separable

(DiCarlo, Zoccolan, and Rust 2012). This is shown in experiments where primates are shown images of objects, and linear readouts of populations of IT neurons allow for a high performance in decoding the class of an object (DiCarlo, Zoccolan, and Rust 2012). Notwithstanding the details of the functional physiology of these high-level areas, the point is that to a first approximation, the ventral stream seems to follow the initial logic proposed by Hubel and Wiesel (and described above for area V2) where at each stage of the ventral stream, neurons acquire larger, more complex, and possibly more invariant receptive fields by integrating over the outputs of previous areas.

2.1.6) Central and peripheral vision:

Another major characteristic of our visual system is that it is foveated: it dedicates a large proportion of its resources to process information from a small central part of the visual field with high precision. We refer to the part of the visual field that has high precision as central or foveal vision, and to the rest of the visual field as peripheral vision. The precise definition of how much of the visual field constitutes central vision varies between studies, and also peripheral vision can be subdivided into near peripheral and far peripheral vision. As a guide, the rod-free fovea (see below for more detail) occupies approximately 1.25° of the visual field (Curcio and Allen 1990).

This property of the visual system is seen in the photoreceptor distribution in the retina (**Figure 6**), where there is a high density of photoreceptors at the center of the retina (the *fovea),* which gradually falls with eccentricity (i.e. distance to the center) (Rosenholtz 2016). Furthermore, not only does the overall density of photoreceptors change as we move from the fovea to the periphery, but also the kinds of photoreceptors: the cones, which are smaller cells responsible for color vision, that require more light to be stimulated, are more concentrated in the fovea; and the rods, which are more relevant to vision in low-light environments but that do not participate in color vision make most of peripheral photoreceptors (Frisby and Stone 2010).

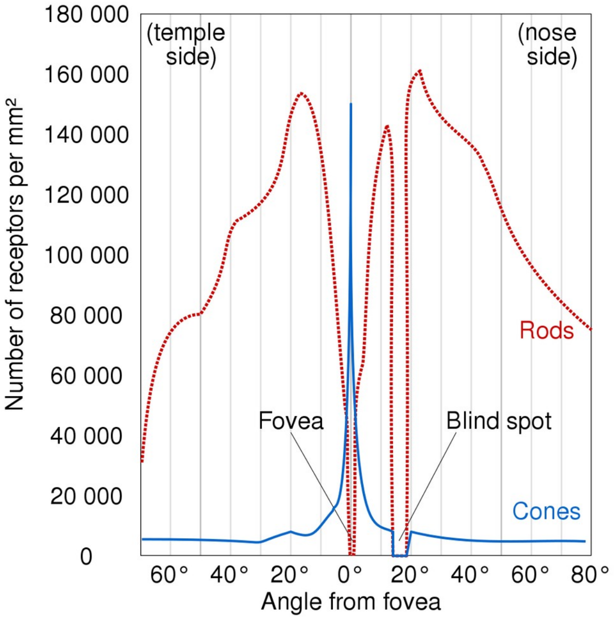


Figure 6. Photoreceptor density across the retina. Plot showing the density of the different photoreceptor cells in the retina at different eccentricities of the visual field. The red line shows the density of rod photoreceptors, and the blue line shows the density of cone photoreceptors. Image from Wikipedia user Cmglee, following (Wandell 1995). CC BY-SA 3.0

This change in cell density is also present for ganglion cells in the retina, which fall from peak densities of 35000 cells/mm^2 near the fovea, to densities smaller than 5000 cells/mm^2 at 4 mm from the fovea, and continue to drop as eccentricity grows (Curcio and Allen 1990). Also, the size of the receptive fields of ganglion cells increases with the distance from the fovea (Peichl and Wässle 1979), so that the integration by ganglion cells in the periphery is done over larger areas, providing coarser representation.

Central vision is also over-represented in the visual cortex. Although as mentioned in **Section 2.1.1**, cortical area V1 represents the visual input in a retinotopic fashion, this mapping of the visual field into cortex V1 does not maintain the spatial proportions of the input. This is measured by an index called the cortical magnification factor, which indicates for each eccentricity the correspondence between distances in the visual field as measured in degrees, and distances in the visual cortex as measured in millimeters. While the magnification factor is large in central vision (i.e. large cortical area per area of visual input), it rapidly decreases with eccentricity, showing a variation of over three orders of magnitude (Van Essen, Newsome, and Maunsell 1984). That is, there are more millimeters of visual cortex (i.e. more cortical resources) dedicated to a given degree of visual field in central vision than in peripheral vision. This cortical magnification factor is roughly proportional to the inverse of retinal eccentricity (Van Essen, Newsome, and Maunsell 1984). This can be seen in **Figure 7**, where a map of the visual scene is shown to the left, and its mapping to macaque monkey V1 is shown to the right (Tootell et al. 1988).

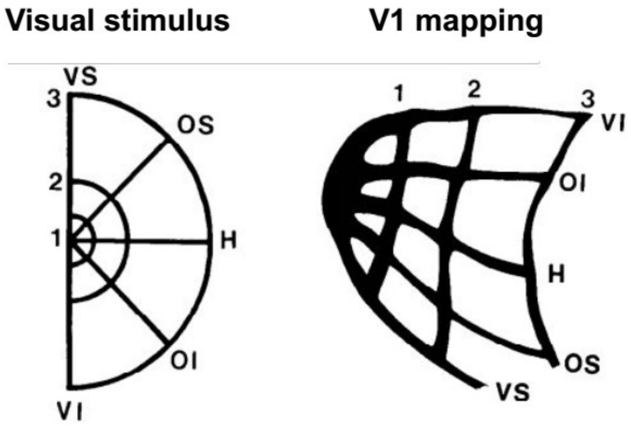


Figure 7. Mapping of the visual scene to cortex V1. To the left, the visual stimulus presented to a macaque monkeys is shown. Before being shown the stimulus, monkeys were injected with radioactive glucose. To the right, the radioactive labeling of cortex V1 corresponding to the lines in the stimulus are shown. As can be seen, there is a large distortion of the dimensions of the visual stimulus when mapped to V1. Adapted from (Tootell et al. 1988).

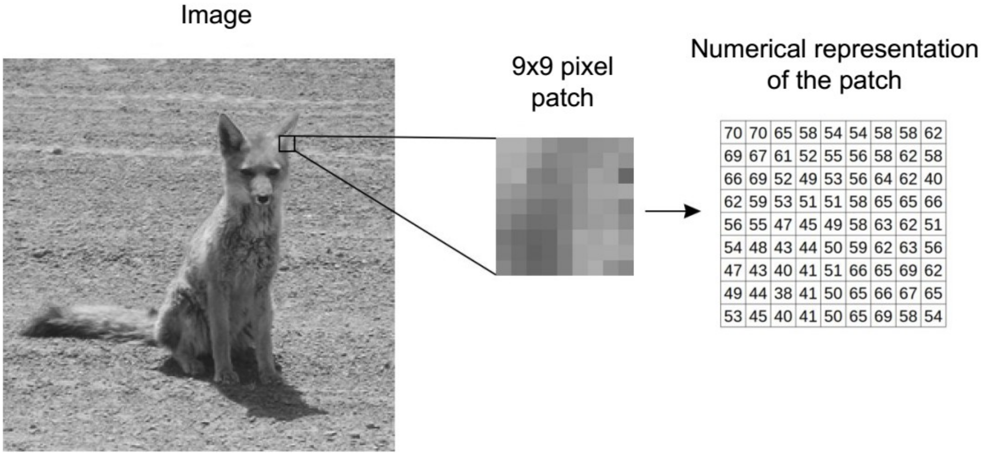
Furthermore, this change in cortical scaling is accompanied by changes in receptive field sizes. In central vision receptive fields of V1 cells are small, and they grow with eccentricity (different functions have been used to relate receptive field size to eccentricity, such as linear functions (Freeman and Simoncelli 2011) and power law functions (Van Essen, Newsome, and Maunsell 1984)). Receptive fields of individual neurons have an average size of around 0.1 deg^2 near the fovea, and they scale to several deg^2 at larger eccentricities (Van Essen, Newsome, and Maunsell 1984). This growth in receptive field size shows that in the periphery, larger areas of the visual input are integrated together by individual neurons. Also, the rate with which receptive field size changes with growing eccentricity varies between areas, with area V2 having a steeper slope than earlier area V1, but with a less steep slope than higher area V4 (Freeman and Simoncelli 2011; Motter 2009).

Finally, it is interesting to note that the visual cortex dedicated to central and peripheral vision may have different connectivity. For example, it is reported that central but not peripheral regions of V1 project directly to V4, and that the parieto-occipital (PO) area only responds to the peripheral visual field (Gattass et al. 2005).

2.2) MODELS OF THE EARLY VISUAL SYSTEM

2.2.1) Visual input as a matrix, and cells as linear filters:

Image computable models of the visual system are models in which the model can receive images as input. In these models, digital images are normally used, which are a 2 dimensional array of numbers in the case of grayscale images (horizontal and vertical dimensions), and a 3 dimensional array when color is included (horizontal, vertical, and color channel dimensions). This way, the value of a given pixel at the *x* horizontal position and the *y* vertical position in a given image can be written down as *I(x,y)* for a grayscale image following the notation in (Hyvärinen, Hurri, and Hoyer 2009). Furthermore, digital images can be thought of as equivalent to the activation pattern of photoreceptors in the retina, with each pixel in the 2D array of the image corresponding to a photoreceptor.



70	70	65	58	54	54	58	58	62
69	67	61	52	55	56	58	62	58
66	69	52	49	53	56	64	62	40
62	59	53	51	51	58	65	65	66
56	55	47	45	49	58	63	62	51
54	48	43	44	50	59	62	63	56
47	43	40	41	51	66	65	69	62
49	44	38	41	50	65	66	67	65
53	45	40	41	50	65	69	58	54

Figure 8. Numerical representation of the visual input. A natural grayscale image is shown to the right, and an amplification of a patch of 9x9 pixels is shown. To the right, the numerical representation of that patch is shown, using a scale of grays going from black (corresponding to 0) to white (corresponding to 1).

If we think of the photoreceptor activation pattern as a digital image, it is natural given our knowledge of the early visual system to model the following early processing steps by the application of linear filters (D. Marr, Ullman, and Brenner 1981; Hubel and Wiesel 1962). A linear filter for a digital image consists of a matrix *W(x,y)*, containing the weights with which the pixels in an image region will be combined to obtain the output of the filter. The output of the linear filter applied at a given point *(x*, y*)* in the image is obtained by centering the filter *W* at this point, and adding together the pointwise products of the filter and the image. For a filter *W* of size (2K+1, 2K+1) this is given by the following formula:

$$\mathrm{O}(\mathrm{x}^{*},\ \mathrm{y}^{*}) = \sum_{x=-K,y=-K}^{x=K,y=K} W(x,y) \times I(x+x*, y+y*)$$

where *O(x*, y*)* is the filter output at point *(x*, y*).* The output of a filter in a given point of an image can be thought of as the response of a neuron which has a receptive field given by the weights in *W*. For example, we can think of the receptive field of a ganglion cell as a filter *W* with positive weights at its center and negative weights in the surround, which in total add to 0. This way, the filter would give an output of 0 to a uniform surface, similar to the response of a ganglion cell, and it would give positive responses to appropriate edges or dots in the underlying image (see **Figure 1**). Furthermore, since ganglion cells tile the retina, we can apply this filter at each point of the image, and thus obtain the whole 2D array of responses of the filter to the image that models the array of ganglion cell responses. The operation of applying a filter to each point of an image is called filtering the image (or convolution of the image). Finally, one last step that could be applied is to introduce a rectification after the linear filtering stage (for example, setting negative values to 0 and leaving positive values unchanged), since using such a linear filter could result in negative outputs which are not biologically plausible.

Then, it is common to model V1 simple cells by using oriented filters similar to the V1 receptive field, where the excitatory and inhibitory regions are stretched in one direction, making the filter selective to oriented structures in images (D. Marr, Ullman, and Brenner 1981; Adelson and Bergen 1985) (**Figure 9**). By filtering an image with different V1-like filters (i.e. with different orientations and spatial frequencies), a feature map can be obtained for each filter, thus representing the activation of different V1 cells in each point.

One popular way of implementing oriented bandpass filters is the Gabor filter, which is built by taking a 2D sinusoidal image, which has a specific orientation and spatial frequency, and multiplying it by a 2D Gaussian function that makes the filter localized in space (R. A. Young 1985; 1986; R. Young, Lesperance, and Meyer 2001; Marĉelja 1980). This kind of image filtering with oriented bandpass (i.e. of a specific spatial frequency) filters is not only important for its relevance to modeling the visual system, it also has important mathematical and practical properties that make them important in engineering applications (R. A. Young 1985; 1986; R. Young, Lesperance, and Meyer 2001). In **Figure 9** we see an example of an image filtered with a set of Gabor filters, with the bottom row showing the filter outputs that represent the activity of a model V1 population.

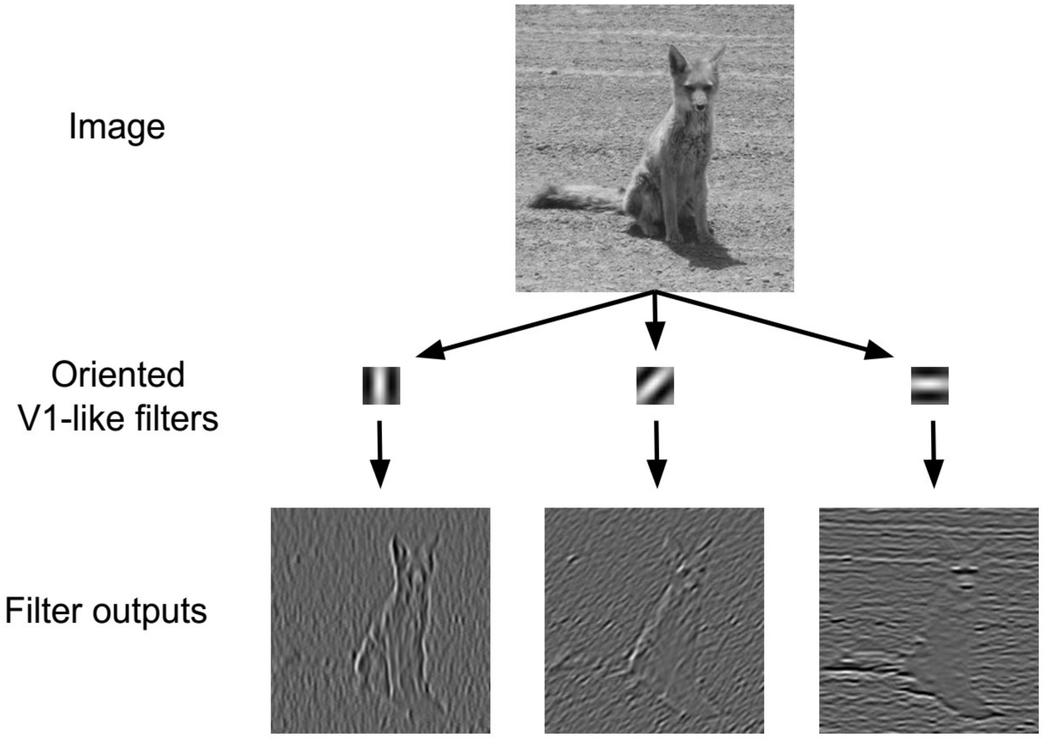


Figure 9. Image filtered with oriented filters. Top row: A natural image. Middle row: The visualizations of three oriented image filters are displayed. Brighter colors represent higher weights in the filter, and darker colors represent smaller weights. The mean value of the filter weights is 0 (thus darker colors show negative weights). Bottom row: The linear outputs of applying the filters to the image above are shown (smaller versions of the filters were actually used, but the images in the

middle row were scaled to aid visualization). Bright pixels indicate higher values of the filter output, and dark pixels indicate smaller values.

2.2.2) Textures, summary statistics and the Portilla-Simoncelli model

Although they do not have an easy clear-cut definition, visual textures are an important element of the visual world. They can be described to a first approximation by appealing to the distinction between “things” and “stuff” (Adelson and Bergen 1991). The former constitute objects, or isolated entities such as lines or blobs. The latter constitutes substance, or the material from which things are made (Landy 2013; Adelson 2001). While objects and shapes in images correspond to “things”, textures and patterns in images corresponds to “stuff”.

While the perception of things, or objects, is often taken to follow from sophisticated processing of the visual input to extract the shape of the thing in the world (David Marr 1982; Riesenhuber and Poggio 1999), processing of texture depends on the patterns produced by the “stuff”, and not so much on their precise layout. Mathematically, visual patterns can be described or encoded using summary-statistics (SS). SS are statistical parameters computed over the pixels of an image, or over the feature maps obtained from an image, and they encode the general structure of a texture without specifying the layout of image features. For example, **Figure 10** shows the output maps of two linear filters applied on an image, and a scatter plot of the pair of outputs at each image location. Although the pattern of filter activations may be complicated, some information of these patterns is encoded in the correlation between the outputs of the two filters across pixels ($\rho = 0.23$), which is a SS of the image, and that is clearly insensitive to the specific distribution of filter activations in the image (i.e. several different feature maps and scatter plots can give rise to the same correlation of 0.23).

But SS are not only useful for modeling and describing texture images, they have also been widely studied as an encoding scheme used by the visual system. In his pioneering work, Bela Julesz generated visual textures, where the values of individual pixels were random, but they were sampled from a probability distribution with specific statistical parameters (Bela Julesz 1962). In his work, different statistical parameters were used, with a parameter of Nth order describing the probability of N-wise point patterns. For example, first-order parameters describe the average intensity (or average pixel value) of a region. Second-order parameters describe the probability of two-pixel patterns, and they are equivalent to pairwise pixel correlations in an image when pixel intensities have a multivariate Gaussian distribution. Third-order parameters mark the co-occurrence of three-pixel patterns, and so on (B. Julesz 1962). He then studied whether humans could preattentively discriminate textures that differed in the Nth order statistics, conjecturing that above some specific order N, textures would be indistinguishable to humans. While initially this line of work showed that textures with identical second-order statistics were not easily discriminable, later work from Julesz and collaborators found examples of textures that were identical up to their third-order statistics and could still be discriminated (B. Julesz, Gilbert, and Victor 1978). Other researchers at the time working with textures used a wide range of texture stimuli, and developed models of different nature from those above, such as models based on the statistics of texture elements (textons) rather than pixel statistics (Beck, Prazdny, and Rosenfeld 1983), making for a vast and varied literature on the topic.

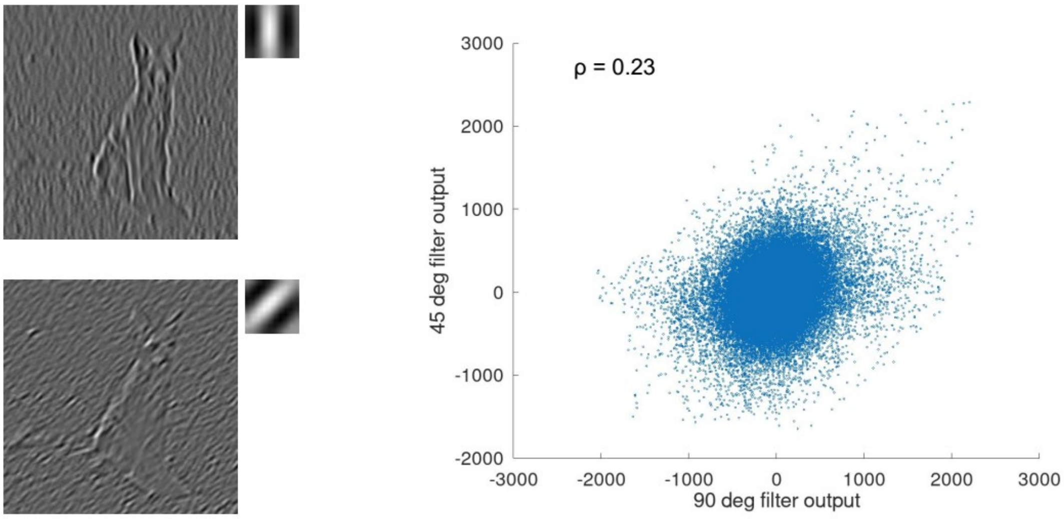


Figure 10. The correlation between filter outputs is a SS. To the left, the filter output maps of two filters are shown. To the right, a scatter plot shows for each pixel the values of the two filter outputs. Thus, the filter outputs are first summarized as a scatter plot (with precise spatial information lost), which is then further summarized into the correlation between the two outputs.

After further development in the study of texture perception and physiology, the relevance of considering that the early visual system preprocesses the visual input with feature extractors (Hubel and Wiesel 1959; 1968; J. Y. Lettvin et al. 1959), and that SS may be computed over the image features started to become evident (B. Julesz et al. 1973). This is the direction that texture perception modeling then took, with the appearance of the filter-rectify-filter model (FRF model) (Bergen and Landy 1991; Bergen and Adelson 1988). The first stage of the FRF model involves filtering an image with a bank of oriented bandpass filters (i.e. V1-like filters), applying a rectification function to the filter outputs (one common rectification function is the squaring of the filter outputs) and sometimes follow these by spatial pooling or other operations as calculating the difference between orthogonal orientations. This linear filtering followed by rectification generates feature energy maps that indicate the “amount” of a feature in a region, which is taken to correspond to the local texture. Finally, a second oriented filtering stage is applied to these non-linear feature maps to find local texture changes that can be used for texture-based segmentation (Bergen and Landy 1991; Landy 2013; Rosenholtz 2014). This family of models was so common in texture modeling that it was also called the “back-pocket model”, and many studies have compared human texture discrimination performance to the outputs of these models (Bergen and Landy 1991; Landy 2013).

Following these lines of research, the description of textures as SS over oriented bandpass filter outputs continued to develop. This progress also involved the elaboration of algorithms that allow to synthesize random textures matched to a set of input SS. One such early model is the Heeger-Bergen algorithm, which uses the histogram of activations of each filter in a filter bank as the texture descriptor, and generates synthetic textures by matching the filter activation histograms of a noise image to a set of reference histograms obtained from an input image (Heeger and Bergen 1995).

A posterior key development for texture modeling and perception research was the development of the Portilla-Simoncelli (PS) algorithm (Portilla and Simoncelli 2000). In this model, several sets of SS are computed over the outputs of a bank of oriented bandpass filters (conceptually similar to **Figure 10**). In this process, the image is first linearly filtered with a pair of quadrature filters for each orientation and scale. The outputs of these are equivalent to V1 simple cell outputs. Then, at each point, the magnitude of the vector given by the activation of the pair of quadrature filters is computed. This magnitude is the energy of the filter at that point in the image, and it throws away the phase information. The output of these energy filters is equivalent to V1 complex cell outputs.

After the linear filter outputs and the energy filter outputs are computed, different sets of statistics are obtained: marginal pixel statistics, linear filter correlations across space and scale, mean energy activations, and energy filter correlations across space, scale and orientation. Note that each of these sets of SS contains several individual SS. For example, the correlations across space involve correlations between several different distances, and the correlations across orientation involve correlations across all the different orientations. How many SS are included in the model depends on the number of orientations and scales used to filter the image, as well as the cutoff neighborhood for computing spatial correlations. Using 4 scales, 4 orientations and a neighborhood size of 7 pixels as in the original work results in a total of 710 parameters in the model. Finally, besides the model of SS used to encode textures, an algorithm allows to iteratively match a given input image (usually noise) to have a prefixed set of values in these SS (Portilla and Simoncelli 2000) (see **Figure 11A**).

Many things are remarkable about the PS model. One is that when we extract the values of these SS from a natural texture image, the new textures synthesized by the algorithm to match these SS have a very appealing and naturalistic appearance, and are often very similar to the original texture (see **Figure 11**). Although in some examples it is evident that the PS model is unable to capture relevant image structure (see **Figure 11B**), the similarity between the original images and the matched synthetic images suggests that the PS model captures important aspects of our texture perception.

Another remarkable feature of the PS model is that it is, in some way, a conceptually straightforward continuation of the feedforward hierarchical cascade by which the cells in a given visual processing stage combine the output of the cells from the previous area into more elaborate receptive fields. For example, the PS correlations across scale could correspond to a processing stage after V1 where the outputs of V1 neurons with different scales are combined. And in fact, it has been observed (as mentioned in **Section 2.1.4**) that neurons of mid-level visual areas, close to V1 in the visual hierarchy, show selectivity for PS statistics (Freeman et al. 2013; Ziemba et al. 2016; Okazawa, Tajima, and Komatsu 2017; 2015). Therefore, although PS statistics are not claimed to be a complete model of human texture perception, or to be fully describe the selectivity of neurons in mid-level visual areas, the PS texture model is an important model of the early stages of visual processing. Besides this described line of work, PS textures have previously been used to probe the changes in selectivity and tolerance when going from area V4 to IT (Rust and DiCarlo 2010).

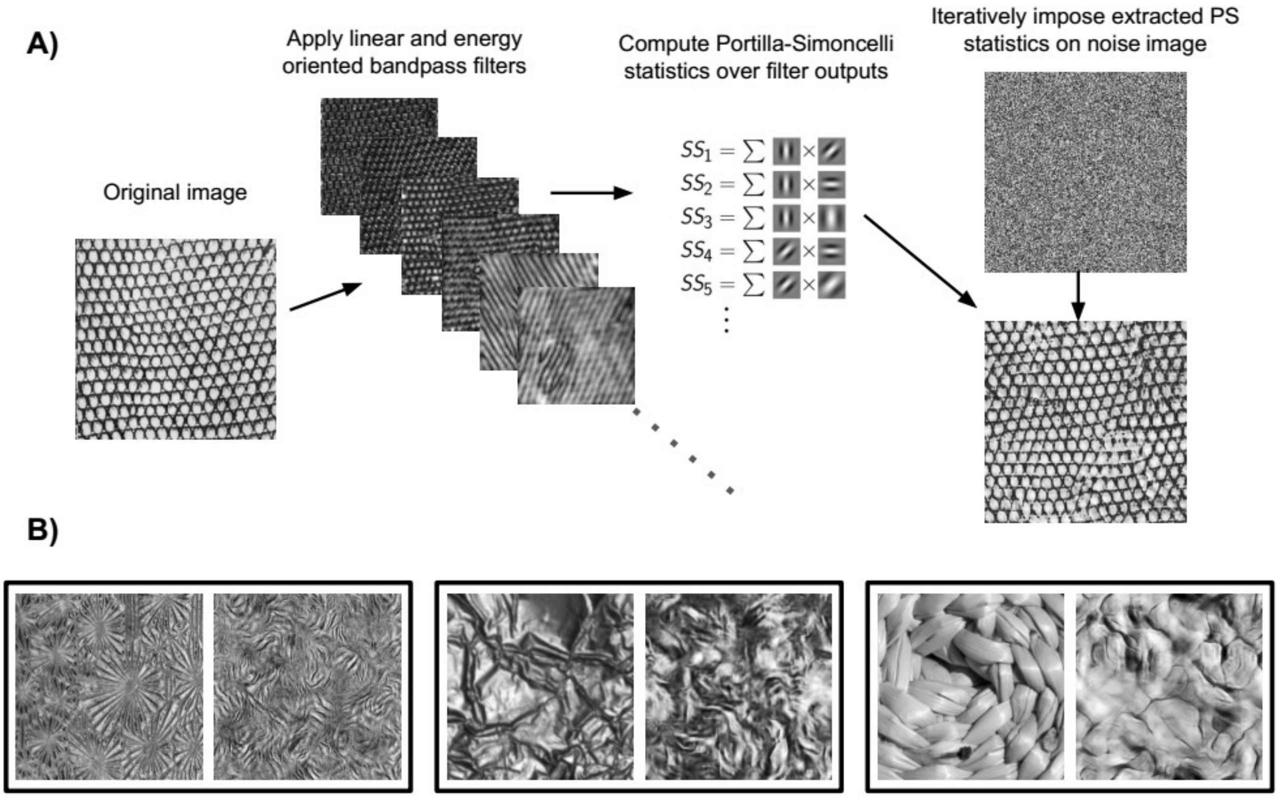

A)
Apply linear and energy oriented bandpass filters
Compute Portilla-Simoncelli statistics over filter outputs
Iteratively impose extracted PS statistics on noise image
Original image
$SS_1 = \sum$ ×
$SS_2 = \sum$ ×
$SS_3 = \sum$ ×
$SS_4 = \sum$ ×
$SS_5 = \sum$ ×
B)

Figure 11. Portilla-Simoncelli texture synthesis. A) Diagram showing the process of synthesizing a PS texture. To the left, an original input image is shown. Then, the output maps of different linear and energy bandpass filters applied to the input image are shown. Third, an illustration of the PS statistics is shown, where each SS involves a correlation between the outputs of two different filters. Finally, a uniform noise image is shown before applying the PS algorithm to match its statistics to the values extracted from the original image, and below it the final synthetic image is shown, for which the SS have been matched to the original. **B)** Three example pairs of original images (left image in each box), and the corresponding PS texture matched in the PS statistics (right image in each box). It can be appreciated that for each image a considerable part of the perceptual qualities of the image are captured, but some levels of information and structure are lost.

Following the aforementioned lines of research, a recent major advancement in texture modeling was the use of features extracted from deep neural networks (DNNs) to compute the SS to describe a texture or an image (L. Gatys, Ecker, and Bethge 2015). In this work, instead of computing SS of co-occurrences of hand-engineered features inspired on visual system physiology, the co-occurrence SS are computed over features of a neural network trained to do object recognition. This model uses features that represent different levels of image structure (i.e. features from earlier layers more similar to V1-like filters, of features from deeper layers that respond to more complex shapes or textures like mid-level and high-level visual areas), leading to the capacity to synthesize considerately more realistic and complex textures.

2.2.3) Fourier representation of images:

The Fourier transform is a mathematical transformation that is frequently used for processing signals such as images, as well as for studying the visual system (De Valois and De Valois 1980). It is based upon the fact that any discrete signal can be represented as a sum of sinusoidal functions, each with its corresponding amplitude and phase (Hyvärinen, Hurri, and Hoyer 2009). Thus, the transform consists on transforming the representation of the signal from the original space, where the value of the signal at each location is specified, to the representation where the signal is specified by the amplitudes and phases of a set of sinusoidal basis functions. For a 2D signal such as an image, this is summarized by the following formula:

$$I(x,y) = \sum_{h=0}^{H-1} \sum_{v=0}^{V-1} A_{h,v} cos(\omega_h x + \omega_v y + \Phi_{h,v})$$

where ω_h and ω_v are the frequencies of the 2D cosine component in the horizontal and vertical directions respectively, $\Phi_{h,v}$ is the phase offset of the cosine function, and $A_{h,v}$ is its amplitude. The terms H and V limiting the bounding the sum are half of the horizontal and vertical dimensions of the image, respectively. The original image can thus be encoded in the set of parameters $A_{h,v}$ (amplitude) and $\Phi_{h,v}$

(phase) for each *h* and *v*. Moreover, each pair or frequencies ω_h and ω_v corresponds to a sinusoidal grating with a spatial frequency ω given by the modulus of the vector (ω_v, ω_v), $\omega = \sqrt{\omega_v^2 + \omega_h^2}$, and an orientation θ given by the direction of the vector, $\theta = \arctan \frac{\omega_v}{\omega_h}$. Therefore, we can also refer to the image parameters by the orientation and spatial frequency of these sinusoidal gratings as $A_{\omega,\theta}$ and $\Phi_{\omega,\theta}$ (Hyvärinen, Hurri, and Hoyer 2009).

Besides its utility for image processing and engineering, the Fourier representation of images is also widely used to study and model visual perception and early visual processing. As mentioned in **Sections 2.1.2, 2.2.1 and 2.2.2**, the neurons of cortex V1 show specificity to both orientation and spatial frequency. The Fourier transform provides a formal way of defining and measuring the content (also *power* or *energy*) of different orientations and spatial frequencies in an image, which are given by the squared amplitudes of the sinusoidals with the corresponding frequency ω and orientation θ, $A^2_{\omega,\theta}$. Also, the Fourier transform allows to separate the energy of the different orientations and spatial frequencies, given by their energies $A^2_{\omega,\theta}$, from their specific layout, given by their phases $\Phi_{\omega,\theta}$. These two different components of the image are called the *Fourier amplitude spectrum* (or power spectrum) of the image, and the *Fourier phase spectrum* respectively.

This distinction between the Fourier amplitude spectrum and the Fourier phase spectrum, and the ability to separate the two is a widely used tool in vision science. Although it is known that most of the scene information is contained in the phase spectrum (see **Figure 12**), the Fourier power spectrum guides important aspects of both our physiological and perceptual responses to images. For example, the Fourier power spectrum is equivalent to second-order pixel statistics, or to pairwise pixel correlations, which were shown by Julesz (B. Julesz 1962) and later by other researchers (Hermundstad et al. 2014; Bergen and Adelson 1988), to be a stronger segmentation cue than the higher-order statistics (HOS) that are contained in the phase-spectrum (although the latter can also produce segmentation (Zavitz and Baker 2014; Barth, Zetzsche, and Rentschler 1998; Beck, Prazdny, and Rosenfeld 1983)).

Also, as described in **Section 2.1.2**, complex cells in V1 are invariant to the phase of the underlying stimulus, unlike simple V1 cells. This is sometimes described as complex cells responding to the local Fourier power content of a given orientation and spatial frequency (Bergen and Adelson 1988; Hyvärinen, Hurri, and Hoyer 2009). Curiously, the mean output response of both linear and energy V1-like filters are approximately given by the spectral content of the image (Hyvärinen, Hurri, and Hoyer 2009; Freeman et al. 2013).

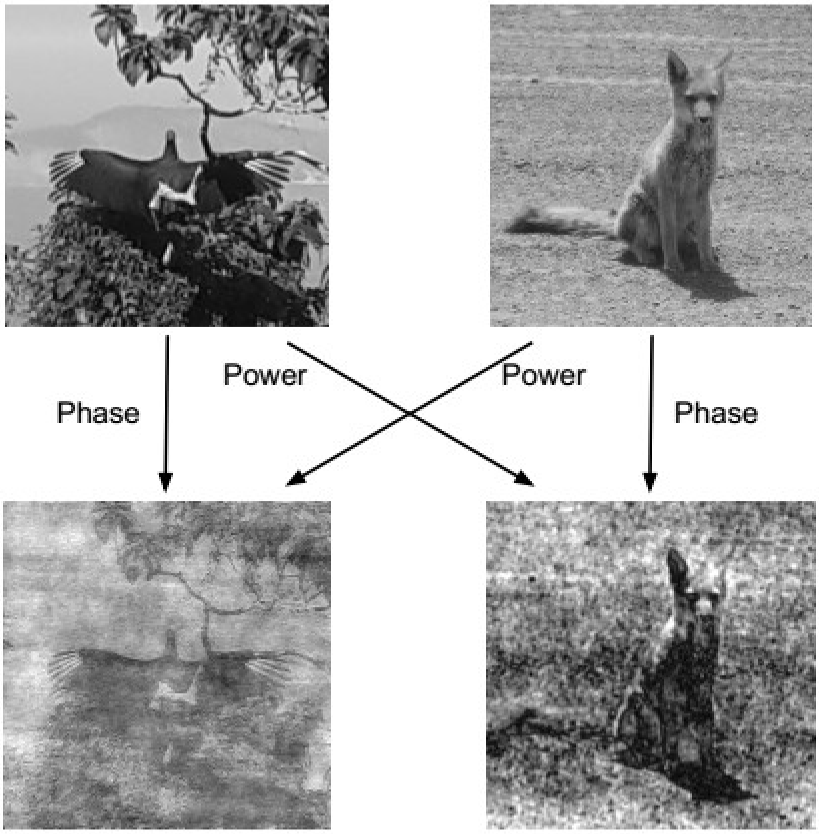


Figure 12. Fourier decomposition of images and phase swapping. The top row shows two original natural images. The Fourier transform was applied to each image, thus obtaining the Fourier power spectrum (i.e. the set of $A_{\omega,\theta}$ for each image) and the Fourier phase spectrum (the set of $\Phi_{\omega,\theta}$ for each image). Then, two hybrid images were generated by combining the Fourier power spectrum of each image with the phase spectrum of the other image. Finally, the new hybrid images generated in the Fourier domain were reconverted to the pixel domain, and the resulting images are shown in the bottom row, with arrows indicating the source of the phase and the power spectrum of each hybrid image. It can be seen that under inspection, the hybrid images are most similar to the image with which they share the phase spectrum.

Another topic in which the Fourier transform has been relevant is in the study of adaptation of the visual system to natural image structure. A well studied property of natural images is that the average power at different frequencies falls roughly as a power law: $A_\omega^2 \propto \omega^{-\beta}$, with $\beta \sim 2.5$ (Thomson and Foster 1997; Tolhurst and Tadmor 2000), and both the spatial frequency selectivity among V1 neurons (Field 1987) and human perception (Tolhurst and Tadmor 2000) seems to be adapted to this property of natural images, making them highly sensitive to the spectral characteristics of the visual input.

Given the sensitivity of early visual processing to the Fourier power spectrum, when performing experiments with visual stimuli, it is a common practice to control for changes in the Fourier power spectrum of the stimuli. For example, matching the power spectrum of two stimuli that are to be compared (as in **Figure 12**) is usually a way to control for the effects of their low-level visual properties (Willenbockel et al. 2010; De Valois and De Valois 1980). Furthermore, when the relevance of "higher-order" features of an image, such as the presence of an object, or scene identity are tested by removing these features, it is common practice to do so by the procedure of *phase-srambling*, where the amplitudes of the Fourier spectrum are kept but the phases are randomized (e.g. (Gong et al. 2018; Freeman et al. 2013)). This phase-scrambling procedure maintains the amount of energy for each orientation and frequency, while destroying their spatial layout. Relatedly, it has been reported that neurons in V1 have the same average response to PS textures, where there is strong higher-order structure, than to their phase-scrambled counterparts (Freeman et al. 2013).

2.2.4) Natural statistics, efficient coding, and contextual interactions:

Encoding the visual input is expensive. It is a complex signal that contains large amounts of useful information. On the other hand, the brain is an organ with finite resources that must be used efficiently. The high cost of representing this visual input and the need to efficiently allocate the finite resources of the brain can be seen, for example, in the foveated structure of our visual system, with a large part of its resources dedicated to processing the very small part of the visual input (**Section**

2.1.6). Simple extrapolation tells us that processing the whole visual field with this high degree of precision would be prohibitive.

When information theory emerged in the mid 20th century, formalizing the concepts behind efficient encoding and transmission for electronic communications devices, it was thus natural that this formalism was borrowed by neuroscience to try to make sense of sensory processing. One of the key concepts of information theory is that of the redundancy in a given signal, or information source, which relates to its predictability and statistical regularities. It was soon realized in the pioneering work of Attneave (Attneave 1954) and Barlow (H. B. Barlow 1961) that the reduction of the redundancies in the input signal, or equivalently efficient coding, may be a guiding principle of biological sensory systems. In fact, these scientists discuss how the visual input we receive is highly redundant, with strong statistical dependencies of visual measurements across space and time. These interdependencies are inherited from the structure of the outside world producing the visual inputs we receive (e.g. such as the co-variation of brightness across an image from a shared light source, or the expected similarity of brightness in neighboring regions of an image which usually belong to the same surface). Using a neural representation of the visual input such that the statistical independence between the responses of the different neurons is reduced may thus increase the coding capacity of the visual system.

Although thinking on efficient coding has much advanced, and it is recognized that a simple reduction of the redundancy of the visual input is likely not a desirable goal for the brain (Barlow 2001), the efficient coding hypothesis and its variants have been an important guiding principle in neuroscience. At least in part, this is because it offers a clear and formal way in which to relate physiological and perceptual phenomena to the statistical structure of visual inputs (Simoncelli and Olshausen 2001). It is possible to propose a criterion of encoding optimality for a representation of images, and then find the representation or code that is optimal for the images of the natural world. These can then be compared to the representations used by the visual system, thus providing a way to compare the guiding theory with visual physiology or perception. For example, different methods have been used to learn a bank of linear filters for representing images, such that their statistical interdependencies are minimized, and these frequently give rise to filters similar to

V1 receptive fields (Olshausen and Field 1996; Bell and Sejnowski 1997). This finding suggests that the selectivity of V1 cells may be explained by its efficiency for encoding natural images under some constraints.

Another important phenomenon that has been explained by appealing to natural image statistics and efficient representations is contextual modulation. For example, early work applying standard statistical techniques to learn efficient non-linear encoding schemes to natural images have shown an emergence of complex contextual modulation phenomena such as end-stopping (Krieger and Zetzsche 1996; C. Zetzsche and Rhrbein 2001; Christoph Zetzsche and Nuding 2005). Other related work optimizing a model of divisive normalization in V1 (i.e. where the response of a neuron is divided by the output of other local cells) to reduce statistical dependencies between V1-like filters, showed contextual modulation behavior similar to that of real V1 cells (Schwartz and Simoncelli 2001). Further models trained to infer statistical dependencies between center and surround taking into account the separation of images into segments (where center and surround could belong to a homogeneous surface or to an heterogeneous surface) could reproduce physiological and perceptual phenomena (Coen-Cagli, Dayan, and Schwartz 2012). These models could also explained variability in contextual modulation across natural images, with images that were inferred to be heterogeneous (or “segmented”) by the model showing significantly smaller contextual modulation than images inferred to be homogeneous (Coen-Cagli, Kohn, and Schwartz 2015).

The analysis of natural image statistics has also explained several perceptual phenomena (Geisler 2008; Burge 2020). One interesting example related to texture perception is the report that human perception of different sets of texture statistics follows their predictability in natural images (Tkacik et al. 2010). In this work, it is shown that humans show low sensitivity to higher-order texture statistics that are predictable from other statistics in natural images (i.e. they carry little information), while showing higher sensitivity to higher-order texture statistics that are unpredictable (they are informative). This finding is also confirmed and extended by later studies showing that the perceptual saliency of different texture statistics follows their relative variability, which would be expected from an efficient coding regime under certain conditions (such as noise levels and channel capacity) that are argued to reign coding in the visual cortex (Hermundstad et al. 2014; Tesileanu et al. 2020).

2.3) PERCEPTION IN PERIPHERAL VISION

2.3.1) Peripheral vision and visual crowding:

As described previously, one of the most salient aspects of the primate visual system is its segregation into a high acuity central vision and a much less detailed peripheral vision. But although this phenomenon is readily evident in everyday perception, it is less clear precisely how peripheral and central vision differ. One first guess may be that peripheral vision is “more blurry”, but it is commonly accepted that this is far from explaining all the properties of peripheral vision (Rosenholtz 2016; Strasburger, Rentschler, and Jüttner 2011). In fact, although the threshold size at which letter can still be identified scales linearly with eccentricity, it does so with a relatively shallow slope, and we thus retain a very decent ability to recognize small letters shown in the periphery (Rosenholtz 2016).

Although there are many characteristics that differentiate central and peripheral vision, some of them probably undiscovered, a phenomenon called visual crowding is commonly considered the main limiting factor of peripheral vision (Rosenholtz 2016; Strasburger, Rentschler, and Jüttner 2011). Visual crowding is a phenomenon in which the recognition or discrimination of a given object in peripheral vision, which would be easy with the object shown in isolation, is impaired by the presence of nearby objects (Whitney and Levi 2011; Rosenholtz 2016). This example of situations in which the acuity of peripheral vision may be sufficient to carry out a recognition task but this task is impaired by contextual interactions comes to show again that acuity loss is not the full picture of peripheral vision limitations.

In a classical experimental paradigm for crowding, a subject is asked to fixate on a point on a screen, and to recognized stimuli shown in the visual periphery (**Figure 13**). Then, different uninformative surrounding stimuli are introduced close to the target, usually inducing a deterioration in task performance (D. G. Pelli, Palomares, and Majaj 2004; Rosenholtz 2016; Whitney and Levi 2011). With this kind of paradigm, crowding has been shown to occur for several kinds of stimuli, and for information at several levels of image complexity, such as vernier stimuli (Levi, Klein, and Aitsebaomo 1985; Manassi, Sayim, and Herzog 2012), letters (D. G. Pelli, Palomares, and Majaj 2004; Strasburger, Rentschler, and Jüttner 2011), faces (Farzin, Rivera, and Whitney 2009; Louie, Bressler, and Whitney 2007; Sun and Balas 2014), objects (Wallace and Tjan 2011), scenes (Gong et al. 2018) and motion (Ikeda, Watanabe, and Cavanagh 2013). Since the visual input we receive in the natural world is also usually cluttered, this makes crowding an important phenomenon for everyday vision (Whitney and Levi 2011; Denis G. Pelli and Tillman 2008).

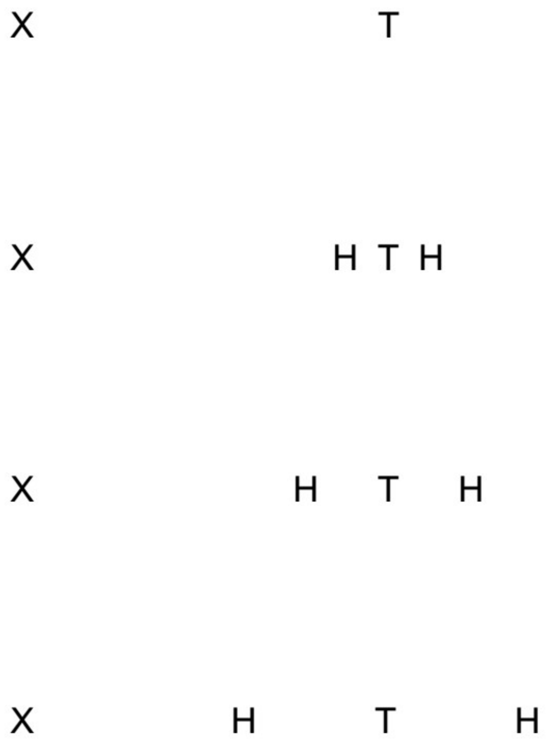


Figure 13. Letter crowding demonstration. Example of classical stimuli configurations. The top row shows an isolated T. Fixating on the cross it can be readily recognizable. On the second row, two flanking letters crowd the T and make it inrecognizable. In the third and fourth row, the crowding effect is alleviated by increased target-flanker distance.

Importantly, crowding has been shown to have many idiosyncratic characteristics. For example, masking is another important perceptual phenomenon in which the detection of a given signal or feature is hindered (or its perceived contrast is lowered) by the presence of surrounding stimuli (Polat and Sagi 1994; Xing and Heeger 2000; D. G. Pelli, Palomares, and Majaj 2004). But unlike masking and other relate phenomena, crowding does not affect target detection, but only target identification (D. G. Pelli, Palomares, and Majaj 2004). This is reflected in the well described subjective appearance of crowded displays, where it is often described that the stimuli are perceived clearly and sharply, but with target and flankers jumbled together into an unrecognizable shape (D. G. Pelli, Palomares, and Majaj 2004; Balas, Nakano, and Rosenholtz 2009). Another major characteristic of visual crowding is that the spatial extent at which flankers interfere with target perception is approximately half of the targets eccentricity (Whitney and Levi 2011; D. G. Pelli, Palomares, and Majaj 2004). This linear scaling of crowding distance with eccentricity is referred to as Bouma's law (D. G. Pelli, Palomares, and Majaj 2004; Bouma 1970). A third important characteristic of crowding observed in many crowding studies is that crowding is asymmetric in the visual field, with more peripheral flankers exerting stronger crowding than flankers more central than the target (Petrov and Meleshkevich 2011; Petrov, Popple, and McKee 2007; Whitney and Levi 2011). Finally, although not specific to crowding, it is known that target-flanker dissimilarity in various dimensions (e.g. color, spatial frequency, kind of stimulus as in letters vs faces, among others) can strongly reduce crowding (Whitney and Levi 2011; Levi 2008).

One reason why crowding research has caught much attention in the last decades is that as a breakdown of object recognition, it is seen as a useful paradigm to study the mechanisms of this computational process. Moreover, although there are many hypotheses about the origins of crowding, probably the most influential is that it happens due to excessive feature integration, which is an otherwise necessary step for object recognition (D. G. Pelli, Palomares, and Majaj 2004; Denis G. Pelli and Tillman 2008; Strasburger, Rentschler, and Jüttner 2011; Levi 2008). This hypothesis posits that, while pooling (or integrating or binding) the set of features that comprise an object is an essential step of object recognition, the pooling windows are too large in the periphery, leading target and surround features to be pooled together, thus preventing recognition of the target, and explaining the “jumbled” appearance of crowded objects (D. G. Pelli, Palomares, and Majaj 2004; Whitney and Levi 2011; Denis G. Pelli and Tillman 2008). This hypothesis is in line with the physiology of the primate visual system, where the receptive fields of neurons grow larger with their eccentricity, thus making for larger integration regions in the periphery than in central vision (**Section 2.1.6**). Furthermore, it has been proposed that Bouma’s crowding window, which grows with eccentricity when measured in the visual field, reflects a pooling window of constant size when measured as area in the visual cortex V1 (Denis G. Pelli and Tillman 2008).

One particularly influential pooling model of crowding is the summary-statistics (SS) representation model of peripheral vision (**Figure 14**). In this model, the step of feature pooling would consist on the computing of the local SS of the features detected in an earlier stage (Balas, Nakano, and Rosenholtz 2009; Rosenholtz 2016; Freeman and Simoncelli 2011). As described for the excessive pooling hypothesis, the regions over which SS are computed would grow with eccentricity in a way that leads to Bouma’s law. This model of crowding has had notable success in many aspects. For example, in a fundamental study in this line of work, it was shown that when computing the SS of the PS texture model for different crowding displays, and then synthesizing texturized versions of these displays, the performance at target classification was similar for the original crowding task in peripheral vision and for careful inspection of these texturized images (Balas, Nakano, and Rosenholtz 2009). Interestingly, this approach reproduced the effects of several variations in display configuration, such as target-flanker dissimilarity. This suggests that the loss of information during crowding may reflect a texture-like encoding of the stimulus in peripheral vision. This kind of encoding scheme could lead to an efficient representation that can sustain behavior with relatively low cost (Balas, Nakano, and Rosenholtz 2009; Whitney and Yamanashi Leib 2018).

Furthermore, this model of crowding has found important parallels in physiology. Another important work on the SS encoding model of peripheral vision tested fixating humans ability to discriminate between two natural scenes that were texturized by applying the PS model locally over a set of windows that grew linearly with the distance from the fixation point. Essentially, subjects were unable to discriminate the two synthetic images when the growth of pooling regions with eccentricity was beneath a critical slope. This was remarkable given that the images were considerably different under visual inspection (i.e. no fixation) (Freeman and Simoncelli 2011), and it was interpreted to reflect a localized texture-like representation in the visual system, such that this representation could not distinguish between the two images when the local SS are matched. Furthermore, the critical slope matched the slope of receptive-field growth of visual area V2, suggesting that the texturized pooling occurred in this area. As described previously (**Sections 2.1.4, 2.2.2**), this was later supported by physiological studies showing that neurons in visual area V2 are selective to these SS and that they show some invariance to changes in texture sample (Freeman et al. 2013; Ziemba et al. 2016). Remarkably, the SS encoding model of peripheral vision has then been used to explain other perceptual phenomena, such as visual search (Rosenholtz et al. 2012), scene perception (Ehinger and Rosenholtz 2016), a wider arrange of crowding displays (Rosenholtz, Yu, and Keshvari 2019), and subjective perception in general (Cohen, Dennett, and Kanwisher 2016).

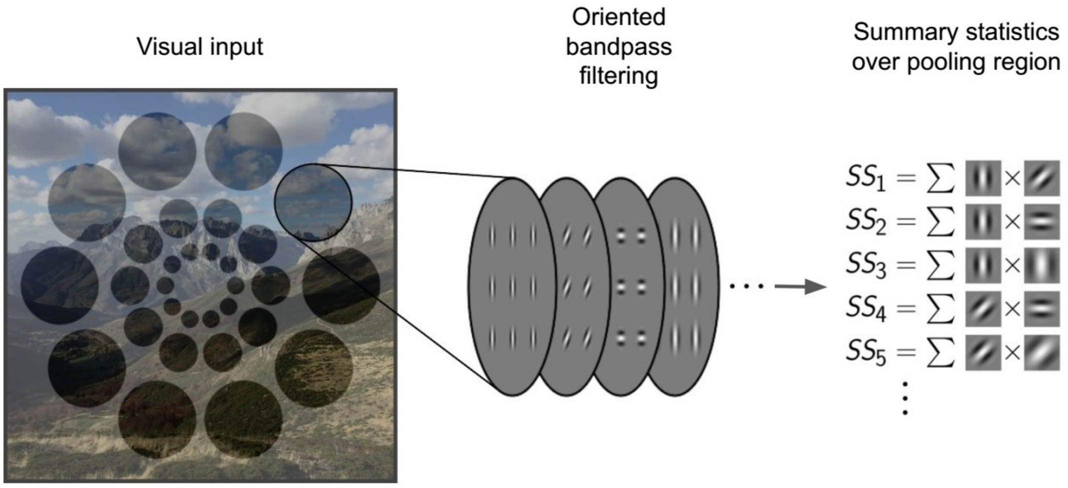


Figure 14. Summary-statistics encoding model of peripheral vision. Diagram showing the steps of the SS encoding model of peripheral vision. The image to the left shows the visual input, and the shaded regions show, for different eccentricities, the receptive fields over which SS are pooled. It can be seen that pooling areas grow with eccentricity. Then, the image is filtered with oriented bandpass filters analogous to V1 cells, and finally the SS over these filter outputs are computed for the pooling region.

2.3.2) Grouping and segmentation in contextual modulation:

Despite the success of the SS encoding model in explaining some experimental results on visual crowding, it has been argued that it fails to explain other important characteristics of crowding, and that this is because of its lack of grouping mechanisms, or of global processing (Herzog et al. 2015) (this is claimed to be a failure of all pooling models of crowding). This is mostly based on several studies showing a strong modulation of crowding phenomena by grouping cues, which seem difficult to reconcile with the feedforward pooling models of crowding (Herzog et al. 2015; Manassi, Sayim, and Herzog 2012; Malania, Herzog, and Westheimer 2007; Saarela and Herzog 2009; Manassi et al. 2015; 2016; Manassi, Sayim, and Herzog 2013). The main type of results that seem difficult to explain with pooling models are: the strong release of crowding induced by subtle manipulations of stimulus configuration that generate target-flanker ungrouping; the release of crowding that can be induced by adding more flankers to a display, if these aid target-flanker ungrouping; and manipulations that modulate crowding by adding flankers far outside Bouma’s window.

Intuitively, under a pooling model where flankers interfere with target identification because their features are pooled together, adding more flankers should worsen performance. Also, subtle changes of the flankers (such as slight rotations, or changes in spacing regularity) that change feature properties only slightly should have small effects on crowding. Finally, flankers that fall outside of the pooling window for the target should have no effect on crowding. Nonetheless, all these manipulations can induce a strong reduction of crowding if they induce target-flanker ungrouping (Herzog et al. 2015). Furthermore, in contrast to (Freeman and Simoncelli 2011) where it was shown that subjects failed to discriminate between two different texturized scenes that shared the local SS, in more recent work (Wallis et al. 2019) showed that subjects can usually discriminate the texturized scenes (generated using either PS or deep neural network-based statistics) from the original scenes, particularly when the latter contained strong non-texture structure (i.e. more “stuff” like content, or stronger grouping cues).

Beyond the particular interest to crowding research, these grouping effects on crowding have been argued to be an important example of the failure of local, feedforward models in general to explain human vision (Herzog and Manassi 2015; Manassi et al. 2016; Doerig et al. 2019). Indeed, different modeling studies found that state of the art feedforward models of the visual system, such as deep neural networks, fail to display these effects of global stimulus configuration in crowding (Doerig et al. 2019; Francis, Manassi, and Herzog 2017; Doerig, Schmittwilken, et al. 2020; Doerig, Bornet, et al. 2020). These studies argue for the need to include recurrent processes of grouping in models of the visual system that allow for global processing of the visual input. In line with this, grouping is known to affect other contextual modulation effects such as backwards contrast masking (Saarela and Herzog 2009), the tilt-illusion (Qiu, Kersten, and Olman 2013), filling-in (Paradiso and Nakayama 1991; Stürzel and Spillmann 2001) and perceptual fading (Vergeer and van Lier 2007), besides also affecting contextual modulation in audition (McWalter and McDermott 2018; Oberfeld and Stahn 2012) and touch (Overvliet and Sayim 2016), indicating that this is an effect with general relevance to perception.

Nonetheless, it has also been argued that some of the mentioned studies may not fully rule out the role of feedforward processing in the observed results. Particularly, it has been noted that our intuitions about pooling models of crowding, which guide most of the experimental designs and analyses mentioned above, may not properly capture the potential behaviors of the SS encoding model (Rosenholtz, Yu, and Keshvari 2019). This is because the SS model proposes a high-dimensional pooling, with several hundreds of statistics being computed over image features, which produces behaviors different from those of the low-dimensional pooling models that usually guide our intuition and toy models. This is illustrated with several examples by Rosenholtz, Yu, and Keshvari (2019), where they use the SS model encoding of experimental crowding stimuli to synthesize new images (referred to as *mongrels*) that show what stimulus information survives and what is lost in this encoding regime. With this approach, the authors show that a number of experimental results which can intuitively seem to show a failure of pooling models, may actually be explained by a high-dimensional SS encoding model. For example, segmentation cues that would intuitively seem to escape such a shallow feedforward model were shown to be conserved in the SS encoding and synthesis approach, and to aid target discrimination. Thus, a feedforward SS encoding model may be able to make use of segmentation cues in the stimuli, without the need for an explicit segmentation process in the encoding stage. Given the success of the SS model in explaining a wide range of phenomena, and its potential for complex unintuitive behavior, it is therefore proposed that stronger and more direct evidence of SS model failure is needed before breaking the parsimony of the model with new ad-hoc mechanisms.

Finally, although several studies have argued for the insufficiency of pooling models, no effort to our knowledge has previously addressed how segmentation and grouping processes may be included explicitly in the SS model.

3) CONTRIBUTIONS OF THIS WORK

The present work has resulted in the elaboration of two published articles, one titled "*Flexible contextual modulation of naturalistic texture perception in peripheral vision*" (Herrera-Esposito, Coen-Cagli, and Gomez-Sena 2021), and another one titled "*Redundancy between spectral and higher-order texture statistics for natural image segmentation*" (Herrera-Esposito, Gómez-Sena, and Coen-Cagli 2021). Here, taking advantage of the broad overview given in the introduction, we summarize the general contributions of these studies.

3.1) Flexible contextual modulation of naturalistic texture perception in peripheral vision

In the first article, we work at the intersection between several of the topics presented in the introduction. As described in the introduction, the most popular model of peripheral vision is the SS encoding model, and its most used implementation is based on the PS model statistics (Rosenholtz 2016). This model has shown close ties to early visual system physiology (Freeman et al. 2013). This is also one of the main models for explaining visual crowding in peripheral vision.

Nonetheless, there is a debate regarding when and why this model fails in presence of segmentation or grouping cues (Herzog and Manassi 2015). In the first article, we bring all of these topics together by doing a psychophysical study of the contextual modulation of PS textures in presence of different segmentation cues.

One problem with previous work studying cases in which the SS model (or pooling models) fail to explain crowding in presence of grouping cues, is that they have used stimuli that are difficult to relate to the SS model. This is because these studies use stimuli that re not naturally defined by their SS, such as very basic geometric shapes (Manassi, Sayim, and Herzog 2013) or complex natural scenes (Wallis et al. 2019), and slight changes to these stimuli may have unpredictable effects on their SS representations (Rosenholtz, Yu, and Keshvari 2019). This has precluded this work from giving a clearer view of what makes the SS model fail. To solve this problem, in this work we constructed our stimuli using PS textures (see **Figure 2** of (Herrera-Esposito, Coen-Cagli, and Gomez-Sena 2021)), which allow us to relate more directly our results to the SS model. We take advantage of this by building a computational observer model that solves the task using a linear readout of the SS of the stimulus, and that we can compare with our experimental results.

Using these stimuli, we showed that geometric segmentation cues led to a reduction of contextual modulation that was not captured by our observer model. Furthermore, we show that for the segmentation cue to reduce contextual modulation, it is important that it induces target-surround discontinuity, and that the low-level properties of the cue and its adjacency to the target seem to contribute little to this effect. The relevance of the high-order property of target-surround continuity, the lack of an effect of the low level properties of the segmentation cue and of its adjacency to the target, and the inability of our observer model to reproduce these results would seem to argue in favor of an explicit segmentation mechanism that modulates the peripheral encoding of target information. This is in support of previous work with simple object-like stimuli (Herzog and Manassi 2015), but generalizing this point to a different kind of stimuli and task that are more directly related to the SS model. We note, however, that the failure of our observer model to capture these phenomena does not exclude the possibility that a more sophisticated SS observer model solving the task (i.e. with several pooling windows, or with a different decoding for solving the task) could reproduce these segmentation results. Future work synthesizing mongrels of our stimuli and testing whether these can reproduce the experimental results can either strengthen the idea that explicit segmentation processes that affect target encoding are required in the SS model (such as flexible pooling windows, or flexible surround suppression), or conversely, show how SS representations can capture quite complex cues that would intuitively seem to escape them, with a novel kind of stimulus.

Furthermore, our stimuli allowed us to test the effect of target-surround dissimilarity at two levels of image structure, each corresponding to a different processing stage of the SS model, and to a different area of the visual system. We separately tested the effect of target-surround dissimilarity in the Fourier amplitude spectrum (or spectral statistics), and dissimilarity in the higher-order statistics (HOS) of the PS model. While the former corresponds to the first stage of the SS model and is associated with area V1, the latter corresponds to the second stage of the SS model and is associated with areas V2/V4. We found that the effect of target-surround dissimilarity was much stronger for the spectral statistics than for the HOS. Furthermore, we showed that this effect was mediated by an increased segmentation between target and surround when the two were dissimilar in their spectral statistics, and not merely by the pooling of target and surround. Although it is well known that spectral statistics are a major segmentation cue, from studies varying texture orientation and spatial frequency or directly in their second-order pixel statistics (Beck 1966; Beck, Sutter, and Ivry 1987; Graham, Sutter, and Venkatesan 1993; Bela Julesz 1962), this is the first study comparing these texture properties to the higher-order statistics of the PS model. This offers important constraints regarding the stages at which contextual modulation and segmentation may occur in the SS model, and it has important implications for understanding the mechanisms underlying target-surround dissimilarity effects in crowding, which could in principle be produced by a pooling mechanisms (Rosenholtz, Yu, and Keshvari 2019).

Also, besides the contribution to better understanding the SS model of peripheral vision, the previous experiment also contributes to better understanding texture segmentation in general. Despite texture segmentation being a widely studied phenomenon, and the PS model being one of the most influential models of texture perception, this is the first experimental test of the HOS of the PS model as segmentation cues. This is important because the few studies that have analyzed the effect of HOS for segmentation have used artificial textures, such as coarse binary textures (Hermundstad et al. 2014), and it was still unclear whether PS statistics play a role in segmentation. This result thus provides an important constraint for the study of texture segmentation psychophysics, physiology and modeling. In the article we provide an extensive discussion of this result in the context of existing models of texture segmentation.

Then, in order to test the relevance of naturalness for our contextual modulation phenomenon, we tested the effect of phase-scrambling the surrounds, which destroys their naturalistic HOS. This procedure has been used previously to show an adaptation of contextual modulation in V1 to the structure of natural images, and has been related to its possible role in the efficient coding of the visual input (e.g. (Coen-Cagli, Kohn, and Schwartz 2015)). We observed that phase-scrambled surrounds generate much weaker contextual modulation than PS textures with naturalistic HOS. Furthermore, we show that this effect of surround scrambling seems to be mediated by increased segmentation. These results support an adaptation of contextual modulation to natural image statistics, and that this adaptation is relevant for human perception. Also, we relate this result to the hypothesis that contextual modulation reflects inference about the center from the structure of the surround (Coen-Cagli, Kohn, and Schwartz 2015). Also, this results suggest a relevant role of V1 surround suppression as an important mechanism limiting texture perception in peripheral vision.

Finally, we tested whether the contextual modulation we observed was due to visual crowding. Remarkably, crowding had not been studied before using textures, to the best of our knowledge. This is important because textures are a major source of information about the environment, and thus the extent to which the phenomenon of visual crowding is a limitation to natural vision will depend on its effects on texture perception. In our experiments we show that our contextual modulation phenomenon does not reliably show one of the hallmarks of crowding: a stronger effect of outwards vs inwards targets. We argue that this may be due to other contextual modulation processes limiting target perception in our task, and that to better understand the limitations of crowding on natural visual perception, it is important to further study how crowding limits different perceptual tasks. Thus, with these results and analysis we argue for the need of some nuance in the widely held view that crowding is the main limitation of peripheral vision.

In sum, in this work we provide an exhaustive analysis of the relation between segmentation and contextual modulation for PS textures. This study contributes to: 1) a better understanding of the limitations of the SS model of peripheral vision, and it provides important constraints to further extend this model, 2) it shows for the first time that the influential HOS of the PS model are a weak segmentation cue, 3) it supports the view that contextual modulation of textures is adapted to natural image structure, and it links our perceptual results to prior physiological literature, and 4) it provides the first analysis of crowding for texture stimuli.

3.2) Redundancy between spectral and higher-order texture statistics for natural image segmentation

In the second article from this work, we look more deeply into one of the results of the first article: the weak role of the HOS of the PS model in texture segmentation. This result is interesting because of the high perceptual relevance of PS statistics, generating an apparent contradiction between the importance of these statistics for the texture perception on the one hand, and their small effect in texture segmentation on the other. To explain this apparent contradiction, we hypothesized that the weak role of these HOS for segmentation can be explained by their redundancy with spectral statistics in natural images for this specific task. In other words, we hypothesized that in natural images the HOS of the PS model would add little information for segmentation over what is already present in spectral statistics, and that resource constraints may push the brain to only rely on spectral statistics for this task. In the second article we explore this hypothesis through a computational study, analyzing the contributions of spectral and HOS to a model observer performing a natural image segmentation task.

In this work we find that although both spectral statistics and the HOS of the PS model are informative for segmentation, using them together improved segmentation performance only weakly. Furthermore, we observed that they tended to mislabel the same images, showing a redundancy between their responses. Nonetheless, we also observed that there are images where HOS do improve segmentation over spectral statistics, which could be exploited by making a “flexible” use of HOS for segmentation, where they are in cases where they would be specially useful. Nonetheless, our attempts to identify these images from their texture statistics had low accuracy, meaning that it may be difficult to implement such a flexible system. This offers a new insight into why these HOS are a weak segmentation cue.

Importantly, although previous work has studied the relation between the saliency of textures in human perception and the natural image statistics (Tesileanu et al. 2020; Hermundstad et al. 2014; Tkacik et al. 2010), these have used artificial textures, and a set of simple HOS. Also, as mentioned, these previous studies looked at the variability of texture statistics in natural images without focusing on any specific perceptual task. In this work we argue that it is important to look at the informativeness of texture statistics (and other image properties) for specific perceptual tasks, in order to understand their role in human vision. This is also in line with the increased interest in developing observer models that solve different tasks taking natural images as inputs, as opposed to observer models using simple tasks and stimuli (Burge 2020). Related to these ideas, we hypothesize that the HOS of the PS model are likely much less redundant with spectral statistics for other tasks such as texture identification or material perception, and that this may explain why these statistics are of such relevance for texture perception but not for texture segmentation. Preliminary data (not shown) comparing the contributions of spectral statistics and HOS to computational observers performing such tasks supports this hypothesis. Thus, this work contributes both insights to the role of an important set of HOS in natural image segmentation, and an example case where the task-specific analysis of natural image structure is used to understand the apparent contradiction between two experimental results.

Finally, an important unanswered question that lingers in our explanation is why, given that both are redundant and similarly useful for segmentation, the visual system would be adapted to use spectral statistics rather than HOS in peripheral vision. That is, the redundancy argument explains why it may be the case that one set of statistics is used and not both together, but it does not weight on which of the two sets of statistics could be preferred. The explanation of why spectral statistics are preferred may be looked for in either physiology or in image statistics.

Possible physiological explanations would propose why it could be beneficial for segmentation to occur using the information represented in V1, the first cortical stage of visual processing. For example, it may be useful to have image segmentation occur as early as possible in the visual system. Also, it may be that the combination of image features represented in V1, which besides local spectral content include color, disparity and motion, make it an ideal location to use all those cues together for segmentation, although selectivity to these features seem to be mostly shared by V2. It may be the case that some particular architectonic feature of V1 (such as specific patterns of horizontal connectivity) subserve a range of functions occurring in this area, including texture segmentation. Maybe the difference in size between the receptive fields of the two areas allows V1 to provide finer segmentation. Or maybe V1 has a smaller number of channels (i.e. fewer types of selectivity) than higher areas V2 and V4, which could make segmentation more efficient.

Alternatively, it may be the case that the explanation to why spectral and not HOS are preferred for texture segmentation in the periphery has an explanation in the computational properties of images. For example, it may be the case that the local spectral statistics can be reliably estimated with smaller areas than the local HOS, which would allow them to sustain finer segmentation. Another example is that a closer look at the patterns of segmentation offered by spectral and HOS shows differences between the two that breaks the symmetric relation of redundancy described in our work. For example, some structures of the world may be particularly important to segment in peripheral vision, and they may be better segmented by spectral than by HOS.

This listing of reasons why spectral statistics may be preferred for segmentation over HOS is not exhaustive. Testing these and other hypothesis constitutes another venue of follow up work on this study.

Flexible contextual modulation of naturalistic texture perception in peripheral vision

Daniel Herrera-Esposito
Laboratorio de Neurociencias, Facultad de Ciencias,
Universidad de la República, Montevideo, Uruguay

Ruben Coen-Cagli*
Department of Systems and Computational Biology and
Dominick P. Purpura Department of Neuroscience,
Albert Einstein College of Medicine, Bronx, NY, USA

Leonel Gomez-Sena*
Laboratorio de Neurociencias, Facultad de Ciencias,
Universidad de la República, Montevideo, Uruguay

Peripheral vision comprises most of our visual field, and is essential in guiding visual behavior. Its characteristic capabilities and limitations, which distinguish it from foveal vision, have been explained by the most influential theory of peripheral vision as the product of representing the visual input using summary statistics. Despite its success, this account may provide a limited understanding of peripheral vision, because it neglects processes of perceptual grouping and segmentation. To test this hypothesis, we studied how contextual modulation, namely the modulation of the perception of a stimulus by its surrounds, interacts with segmentation in human peripheral vision. We used naturalistic textures, which are directly related to summary-statistics representations. We show that segmentation cues affect contextual modulation, and that this is not captured by our implementation of the summary-statistics model. We then characterize the effects of different texture statistics on contextual modulation, providing guidance for extending the model, as well as for probing neural mechanisms of peripheral vision.

Introduction

Central and peripheral vision fulfill different roles in visual perception, as reflected by their different information processing capabilities. The most influential model of peripheral visual processing is the summary statistics (SS) model (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Rosenholtz, 2016), which proposes that the peripheral visual input is represented using SS of the activations of feature detectors (Figure 1), computed over prespecified regions of the visual field (termed pooling windows) whose size scales linearly with eccentricity. This model fits in the descriptive paradigm of vision as a hierarchical feedforward cascade of visual feature detectors (Riesenhuber & Poggio, 1999; Doerig et al., 2019), and it is theoretically appealing because replacing a detailed representation of the visual input with a SS results in a significant compression of the visual input. Furthermore, this compression results in a loss of information that could parsimoniously explain the limitations of peripheral vision (Rosenholtz, 2016), including the impairment of target identification by surrounding stimuli (visual crowding (Balas et al., 2009), often regarded as the most important factor in peripheral vision), as well as phenomena related to visual search (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012), scene perception (Freeman & Simoncelli, 2011; Ehinger & Rosenholtz, 2016), and subjective aspects of visual experience (Cohen, Dennett, & Kanwisher, 2016). The SS framework has also been used to explain auditory perception of sound texture (McDermott & Simoncelli, 2011), suggesting a more general role of SS representations.

Despite providing a solid foundation, it has been hypothesized that phenomena involving segmentation and grouping in peripheral vision escape the standard SS model, and therefore more accurate models of peripheral vision should include recurrent processes of grouping and segmentation (Manassi, Sayim, & Herzog, 2013; Manassi, Lonchampt, Clarke, & Herzog, 2016; Doerig et al., 2019). Grouping different elements into objects or ensembles, or conversely segmenting the scene into different segments, is an essential aspect of human vision. Segmentation processes have been shown to affect several contextual modulation phenomena (i.e. phenomena in which perception of an image region is affected by its surrounds), such as backward contrast masking (Saarela & Herzog, 2009), the tilt-illusion

Citation: 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, 1–26, https://doi.org/10.1167/jov.21.1.1.

https://doi.org/10.1167/jov.21.1.1 Received February 21, 2020; published January 4, 2021 ISSN 1534-7362 Copyright 2021 The Authors

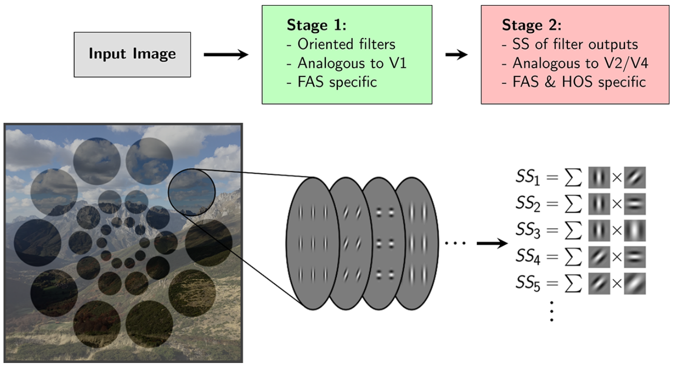


Figure 1. **Summary-statistics representation model.** Illustration of the main features of the standard SS model, and its relation to physiology and image properties. An input image is first filtered with a bank of oriented V1-like filters, whose activation power is determined by the Fourier amplitude spectrum (FAS) of the image in the pooling region. Then SS are computed over the activations of these filters in fixed pooling windows that tile the visual field. The SS in the second stage are referred to as higher-order statistics (HOS; in contrast to the statistics contained in the FAS).

(Qiu, Kersten, & Olman, 2013), filling-in (Paradiso & Nakayama, 1991; Stürzel & Spillmann, 2001), perceptual fading (Vergeer & van Lier, 2007) and crowding (see Herzog, Sayim, Chicherov, & Manassi, 2015 for a review). Similar effects have been reported in audition (Oberfeld & Stahn, 2012; McWalter & McDermott, 2018) and touch (Overvliet & Sayim, 2016). In particular, much work with vernier and letter stimuli showed that even small changes to the contextual stimuli, or changes far away from the target, can lead to target-surround ungrouping and a considerable reduction in crowding (Kooi, Toet, Tripathy, & Levi, 1994; Saarela, Sayim, Westheimer, & Herzog, 2009; Manassi, Sayim, & Herzog, 2012; Manassi et al., 2013; Manassi, Hermens, Francis, & Herzog, 2015; Manassi et al., 2016), a phenomenon known as "uncrowding." It has been argued that these results show a failure of feedforward pooling models, such as the SS model, and that this failure is due to their lack of recurrent processes of grouping and segmentation (Herzog et al., 2015; Francis, Manassi, & Herzog, 2017; Doerig et al., 2019; Doerig, Bornet, Choung, & Herzog, 2020). Furthermore, current SS model implementations also fail to capture the peripheral appearance of natural scenes that contain strong grouping and segmentation cues (Wallis et al., 2019). However, it has been proposed that the SS model may be able to account for these results, without recurrent segmentation or grouping mechanisms that modify the encoding of SS, because segmentation cues could be directly decoded from the fixed SS representation (Rosenholtz, Yu, & Keshvari, 2019). One challenge in exploring these alternatives is that commonly used crowding tasks, such as discriminating the offset of a crowded vernier stimulus (Manassi et al., 2013; Doerig et al., 2019), or the more recent task of discriminating complex scene distortions (Wallis et al., 2019) depend on perceiving a given feature from a specific target object in an array, or complex arrangements of features, which are not easy to link intuitively or computationally to the more distributed and texture-like representations of the SS model (Rosenholtz et al., 2019).

Here, we test more directly the hypothesis that the SS model does not fully capture segmentation effects on contextual modulation, using naturalistic visual textures, which are more easily linked to SS representations. SS representations have long been studied in relation to texture perception, because textures are statistically defined stimuli to texture perception (Julesz, 1962; Julesz & Caelli, 1979; Victor, 1994; the SS model is also referred to as the texture-tiling model of vision; Doerig et al., 2019). We use naturalistic Portilla-Simoncelli (PS) textures (Portilla & Simoncelli, 2000), which have been instrumental to the recent success of the SS model (Balas et al., 2009; Freeman & Simoncelli, 2011; Rosenholtz et al., 2012; Ehinger & Rosenholtz, 2016) and are a useful experimental tool for probing the model. PS textures are defined by a set of SS that are inspired in natural image statistics and early human vision, and which are the basis of the main implementation of the SS model of peripheral vision. This makes it possible to compare directly perception of PS textures to SS model predictions. Furthermore, it has been shown that, different from primary visual cortex (V1), neurons in higher cortical areas V2 and V4 are selective for PS statistics (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Okazawa, Tajima, & Komatsu, 2015; Ziemba, Freeman, Movshon, & Simoncelli, 2016; Okazawa, Tajima, & Komatsu, 2017), offering a framework to relate the SS model and peripheral vision to neural mechanisms. However, no studies have addressed how peripheral naturalistic texture perception is affected by contextual modulation and by segmentation cues (see Meinecke and Kehrer, 1994; Morikawa, 2000; Schade and Meinecke, 2009; Schade and Meinecke, 2011; Victor, Thengone, & Conte, 2013 for examples with artificial stimuli, and Wallis & Bex, 2012 for a study with natural images that does not explore segmentation).

Therefore, we use a PS texture discrimination task to study contextual modulation and segmentation in peripheral vision within the framework of the SS model. We evaluate how different texture surrounds affect texture perception, and study the influences of grouping and segmentation cues and of surround structure, as well as the relation between this contextual modulation and crowding.

Our results reveal an important role of segmentation processes in peripheral perception of naturalistic texture and highlight limitations of the feedforward framework of visual processing. Furthermore, we link our results to existing versions of the SS model and to previous work on the physiology of the early visual system, pointing to possible computational processes that may underlie the results. Our work can provide guidance for implementing and testing extensions of the standard SS model that include segmentation and grouping.

Methods

Participants

A total of 98 adult individual participants (including the authors D.H. and L.G., denoted in the figures by colors blue and green, respectively), participated in the experiments, of which 34 were women. All participants had normal or corrected to normal vision.

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Research Ethics Committee of the Faculty of Psychology of the Universidad de la República. Participants gave signed consent to participate in the experiment, and to have the anonymized data from the experiments made available online. Participants were given no economic or course credit reward for their participation in the experiment.

Texture synthesis

We synthesized grayscale naturalistic textures using the PS texture synthesis algorithm (Portilla & Simoncelli, 2000) in Octave (Eaton, Bateman, Hauberg, & Wehbring, 2015). The algorithm first computes a set of statistics over an input image, including mean luminance, contrast, and higher-order moments of the pixel histogram; and the means and pairwise correlations of the activations of multiscale, multi-orientation filters (steerable pyramid Simoncelli, Freeman, Adelson, & Heeger, 1992) analogous to V1 cells. Then it iteratively modifies a white noise image until its statistics match those of the input image. We used as input images natural textures from the Brodatz texture database, the Amsterdam Library of Textures (Burghouts & Geusebroek, 2009) and from the database presented in Lazebnik, Schmid, & Ponce, 2005. We refer to an image synthesized this way as a naturalistic texture or PS texture. We used filters with four scales and four orientations, and nine by nine pixels neighborhood (corresponding to a 0.3 degrees × 0.3 degrees neighborhood with the viewing distance used) for computing the spatial correlations of the filter responses. We synthesized two 1024 × 1024 PS textures for each input image.

For each PS texture, we also synthesized a phase-scrambled texture. This was achieved by first generating a uniform noise image and then replacing its Fourier amplitude spectrum (FAS) for the FAS of the naturalistic texture. Thus, this procedure produces a pair of PS textures and a pair of phase-scrambled textures that are used in the experiments.

Phase-scrambling a naturalistic image can change the histogram of pixel activations (e.g. changing the minimum and maximum intensities). To prevent participants from using aspects of the pixel histogram (e.g. brightness) as cues to solve the task, we matched the pixel histograms of the naturalistic and phase-scrambled images to an average of the two, using the SHINE package for Octave (Willenbockel et al., 2010) with 30 iterations. In each iteration, their FAS was also matched to the original FAS, and the structural similarity index (SSIM) with respect to the original image was also optimized in order to reduce alterations to image structure (Wang, Bovik, Sheikh, & Simoncelli, 2004; Willenbockel et al., 2010). Images produced by this method appeared very similar to the starting textures (besides changes in pixel intensities), suggesting it did not produce noticeable structural alterations.

In experiment 3, to generate the surround image that was dissimilar to the target only in higher order statistics (HOS), we started by generating a new PS texture using a different input image than the one used for the target. Then we matched its FAS and pixel histogram to those of the target PS texture with the SHINE package, using 30 iterations. In each iteration, the SSIM with respect to the original surround PS texture was also optimized. For the surround texture that was dissimilar in both FAS and HOS, the same procedure was used but without matching the FAS to the target PS texture.

Texture selection

Because there is considerable variation in the discriminability of different PS textures from their phase-scrambled counterparts (Freeman et al., 2013), we synthesized a large set of pairs of PS and phase-scrambled textures and selected those that subjectively appeared to have high discriminability, to make the task easier. We also selected textures that had different kinds of structures, in order to better probe the texture space (e.g. strongly oriented, weakly oriented, regular, and irregular).

In addition, in experiment 3, most textures to which we applied the FAS matching procedure acquired a phase-scrambled appearance, so we selected for further use those that maintained a naturalistic appearance after this procedure.

Due to resource constraints and design choices, we did not use the same number of textures for each experiment. The textures used in each experiment are those shown in the corresponding figure.

Organization of experimental sessions

An experimental session consists of one participant performing an experiment with a given texture. When a participant performed an experiment with more than one texture, these were used separately in different experimental sessions. Participants were allowed to perform as many experimental sessions as they were willing to complete. Nonetheless, no participant performed more than two experimental sessions in the same day, and no participant performed the same experiment twice with the same texture. Excluding the main author, who completed 19 experimental sessions, the rest of the participants completed between 1 and 5 sessions, with a mean of 1.7 and a median of 2 experimental sessions completed by participants. For each experiment, we report the total number of experimental sessions, corresponding to the sum of the experimental sessions performed by all participants. In total, participants completed 189 experimental sessions across all experiments.

Experiment sessions were divided into 2 to 4 experimental blocks (with balanced conditions) separated by 30 seconds resting periods. For experiments 2 and 5, which involve manipulations of stimulus configuration, we also separated the main experimental conditions of these experiments (i.e. target shape for experiment 2 and surround position for experiment 5) into 2 condition blocks that were nested within the main experimental blocks, to prevent possible confusion. Finally, the different conditions contained within a block were randomly interleaved, and each was presented an equal number of times. The total duration of the experiments, including training and instructions, was between 20 and 45 minutes.

Detailed anonymized information on which participants performed each experiment and with which textures can be found in the online data made available for this article (see Data Availability below).

Stimulus sampling

All textures shown in the experiments were patches cropped from these larger synthesized images, with a linear transparency gradient at their border, allowing for a smooth fading with their neighboring surface (e.g. the background or a neighboring texture). These gradients had a length of 4 pixels, roughly equivalent to 0.15 degrees. For each texture patch displayed, the cropped region was randomly selected over the whole image on a trial-by-trial basis.

We note that because the PS statistics were matched over the large synthesized images, the random sampling of patches from these images introduced some trial-by-trial variation in the texture statistics displayed. Although testing the effect of this image variability on our results would require additional experiments, we think this variability is unlikely to have significant effects on the participants' performance, as discussed in section S6.

In each individual trial, an angle multiple of 90 degrees was randomly chosen and all textures were rotated by this angle before being cropped for display. This was done to reduce participants' adaptation to low level properties of the textures.

Task

Our task is a variation of that described by Freeman et al. (2013), and consists in discriminating between the naturalistic and the phase-scrambled versions of a texture.

The target stimuli (targets) consisted of 2 circular patches of texture presented simultaneously for 233 ms, centered at 12 degrees to the right and to the left of the fixation point (Figure 2). We used three different target configurations: (1) phase-scrambled target to the right (PS texture to the left), (2) phase-scrambled target to the left (PS texture to the right), or (3) no phase-scrambled target (PS texture in both targets). The three configurations were shown an equal number of times, in random order. Participants were instructed to report the location of the phase-scrambled target with the arrow keys, and to use the upward arrow to indicate the absence of phase-scrambled targets. This task design with two targets and three conditions was used to discourage participants from looking away from fixation, to compensate for the lack of eye-tracking in the experiments.

The sequence of events in any given trial was the following (see Figure 2): (1) start with the gray screen, (2) after 500 ms, a red fixation dot appeared at the center of the screen, on which participants were instructed to fixate, (3) after a time interval sampled uniformly from 400 ms to 600 ms, the 2 targets were presented simultaneously for 233 ms (14 frames), (4) after the targets disappeared, the participant responded (without a time limit), (5) auditory feedback was provided and the fixation dot disappeared, returning to step 1). Participants were told to use the response stage (step 4) to rest as needed by delaying the response.

For experiment 4, we slightly modified the task for half of the participants. In this variation of the task, participants were instructed to indicate the position of the PS texture, instead of the phase-scrambled texture. Accordingly, we substituted the condition with two naturalistic targets for a condition with two

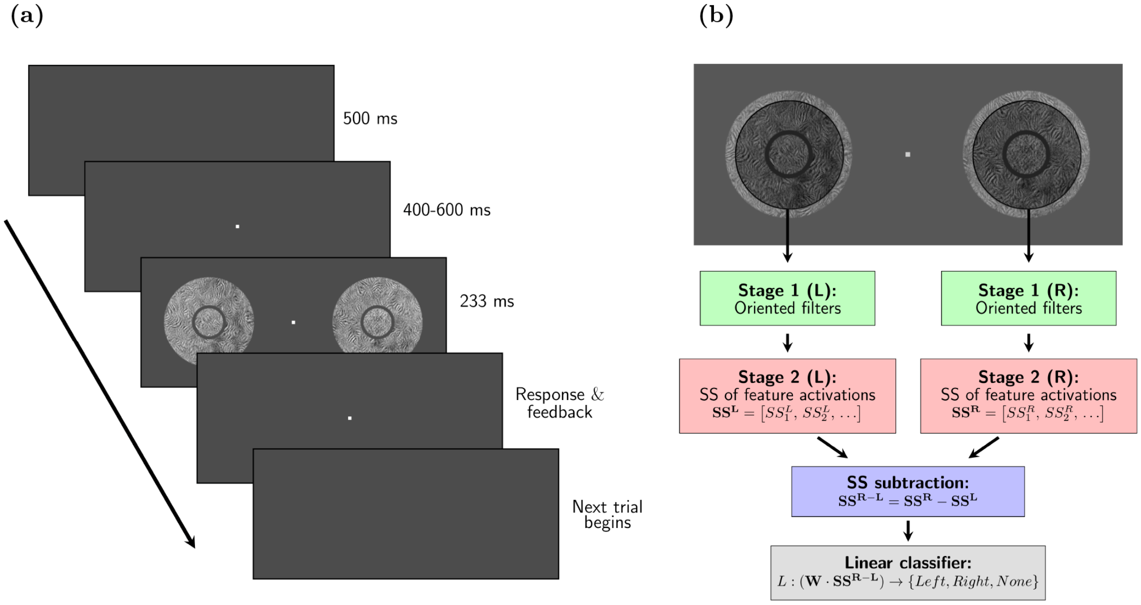

(a)
500 ms
400-600 ms
233 ms
Response & feedback
Next trial begins
(b)
Stage 1 (L): Oriented filters
Stage 1 (R): Oriented filters
Stage 2 (L): SS of feature activations $\mathbf{SS^L} = [SS_1^L, SS_2^L, \ldots]$
Stage 2 (R): SS of feature activations $\mathbf{SS^R} = [SS_1^R, SS_2^R, \ldots]$
SS subtraction: $\mathbf{SS^{R-L}} = \mathbf{SS^R} - \mathbf{SS^L}$
Linear classifier: $L : (\mathbf{W} \cdot \mathbf{SS^{R-L}}) \rightarrow \{Left, Right, None\}$

Figure 2. **Task design and observer model.** (**a**) Two targets centered at 12 degrees to the left and to the right of the fixation point were displayed simultaneously in each trial for 233 ms. Either the left, the right, or none of the targets was sampled from the phase-scrambled texture (with the others sampled from the naturalistic texture), and the participant had to indicate with the arrows where (if) the scrambled texture was present (3 AFC). In most trials, we added uninformative surround textures around the target (in this example, surround textures are present, separated from the target by a gap). In any given trial, the two targets always had the same kind of surround. To aid visibility, the size and color of the fixation dot in this image are not the same as in the experiments. (**b**) Diagram showing the architecture of the model observers based on the SS model used to simulate the experiments. The SS of the PS model are computed over circular pooling windows centered on each target (illustrated by the shaded regions). The difference between the SS of the two targets is used to predict the stimulus configuration (i.e. where the phase scrambled target texture is). See Methods for implementation details.

phase-scrambled targets, thus maintaining the structure of the task.

Surround textures

In all the experiments, we included surrounding textures with varying shapes and texture contents. In any given trial, the two targets shared the same kind of surround. These surrounds were also sampled randomly (and independently from each other and from the targets) from the larger synthesized textures. Unless indicated otherwise in the text, the surrounds were sampled from the PS texture of the texture pair to be discriminated in the targets.

In most cases, surrounds were rings (or half-rings) with a width (i.e. distance between inner and outer edges) equal to target diameter. Experiment 1 and texture T1 in experiment 3 were an exception, having a surround width 1.4 times the diameter of the target. The surrounds of the split disk targets in experiment 2 were not rings, but they had the same outer diameter as the surrounds for the corresponding disk-shaped targets.

Surrounds could be contiguous to the target or separated by a gap showing the gray background. The gap had a width of 0.5 degrees in all cases except experiment 1, where it had a width of 0.35 degrees and texture T1 in experiment 3, where both gaps of 0.5 degrees and 1 degree were used (although these were grouped together for the analysis, see Supplementary section S4). We selected this gap width by subjective visual inspection, considering the need for a gap large enough to be clearly visible in the periphery, but as small as possible to minimize the spatial differences between the stimuli with and without a gap (see Supplementary sections S2 and S4 for more information on the slight variability in gap size in some conditions).

Training and difficulty adjustment

Before the experiment, participants were provided with training opportunity. Auditory feedback was used in all stages of training, as well as in the main experiment. In the first training session, targets were shown without surround and remained on the screen until the participant responded. The second training session also used targets only, but had the same dynamics as the experiment. Both sessions were terminated at will by the participant by pressing a special key.

After running experiments 1 and 3 with texture T1 with a target diameter of 3.5 degrees, we observed considerable variability between participants in task performance. Therefore, we adjusted task difficulty to each participant (except when noted otherwise), in order to drive participants to a more informative performance range (preventing saturation with very high or with chance-level performances). To this aim, we presented a sequence of trials with unsurrounded targets in which target diameter was adaptively adjusted using the accelerated stochastic approximation procedure (Treutwein, 1995) to drive participant performance to a predetermined level of 90% correct responses (see Supplementary section S7 for details on the procedure and final size distributions). If the final target diameter was larger than 5.3 degrees (160 pixels), we used a diameter of 5.3 degrees in the experiments. The widths of the surrounds were then set equal to the target diameter. We note that the results from experiment 1 and experiment 3 with texture T1 were obtained without size adjustment, because this procedure was only incorporated after these experiments.

After size adjustment, we repeated the static and dynamic training stages as described above, including also the surrounds, and instructed participants to perform the task ignoring the surrounds. Again, participants terminated these sessions at will.

Materials and apparatus

The task was performed in a dark room, using a 27 inch LCD screen (ASUS, model PG278QR) with a refresh rate of 60 Hz. Participants used a chinrest to maintain a viewing distance of 40 cm, at which 1 degree of the visual field subtended 30 pixels. Experiments were run on Psychtoolbox-3 (Kleiner et al., 2007) running in Octave version 4.0.0 in Ubuntu 14.04.

The background gray had a luminance of 8.7 cd m^{-2}, and the textures used in the experiments had a range of mean luminance of 50.9 cd m^{-2} to 67.3 cd m^{-2}, and a range of standard deviations in the luminance of the pixels of 26.9 cd $\mathrm{m}^{-2}$to 34.1 cd m^{-2}, as determined with a screen calibration performed with a colorimeter (Cambridge Research Systems, model ColorCAL II).

Summary-statistics model observer

We implemented an image-computable observer model based on the feedforward SS model with fixed pooling windows (Freeman et al., 2013). This model first computes PS statistics over the two stimuli, then computes their difference and feeds it to a linear classifier to solve the task (see Figure 2b). The weights of the discriminator were optimized to maximize discrimination performance on a training set, and the model is then tested on a separate test set (cross-validation). We added noise to the PS statistics computed by the model in both training and testing stages, to roughly match the performance of the human participants on average across stimuli.

We first generated sample images of single stimuli, such as those used in the experiments, with either phase-scrambled or naturalistic targets diameter 110 pixels (corresponding to a diameter of 3.7 degrees in the experiments), and with the different surrounds. We adapted the code of Freeman and Simoncelli, 2011, to compute PS statistics over a circular fixed pooling area centered on the target. We used a pooling area with a diameter of 360 pixels, equivalent to 12 degrees of visual field. We based this pooling size on Bouma's law of crowding (Whitney & Levi, 2011), which says that surround elements hinder target perception when they are within a distance of about 0.5 times the eccentricity, thus we used this distance (12 degrees × 0.5) as the radius of integration around the target center. We note that previous studies on the SS model (Freeman & Simoncelli, 2011; Doerig et al., 2019; Rosenholtz et al., 2019; Wallis et al., 2019) used multiple pooling regions with smaller sizes (with their diameter and not their radius equal to half the eccentricity, analogous to V2 receptive fields) that tile the visual field. Although such models are more realistic than our model, and their structure may allow them to capture some more complex phenomena, using multiple pooling regions would require a more complex decoder and several additional design choices. Therefore, in the interest of simplicity, we opted for the single pooling window matching Bouma's law.

We computed PS statistics using 4 scales, 4 orientations, and a neighborhood for computing spatial correlations of 7 pixels (smaller than for texture synthesis to reduce the number of model parameters), corresponding to 0.7 degrees of visual field in the experiments. This procedure leads to 782 SS per stimulus (after removing the repetitions of symmetric parameters from the correlation matrices).

To mimic the experimental task, we arranged the stimuli (which either had naturalistic or phase scrambled target) into three kinds of ordered pairs, equivalent to those shown in the experiment. Using *Nat* and *Scr* to refer to stimuli with naturalistic and scrambled targets respectively, the three kinds of ordered pairs were $\{Scr, Nat\}$, $\{Nat, Scr\}$, or $\{Nat, Nat\}$. As in the experiment, the stimuli from a given pair had the same surround. Then, we subtracted the SS of the second stimulus to each corresponding SS of the first stimulus, resulting in 782 differences in SS (or predictors) for each stimulus pair. The observer consisted of a linear discriminator trained to predict the class of the stimulus pair (e.g. $\{Scr, Nat\}$, $\{Nat, Scr\}$, or $\{Nat, Nat\}$) from the SS difference of the pair.

First, for an observer trained for a given experiment, we generated 750 stimulus pairs (250 of each class), or trials, for each different surround condition in the experiment, and computed the difference in SS (predictors) for each generated pair. We then added Gaussian noise to the predictors, with a standard deviation equal to the standard deviation of the predictor across the training dataset containing all the conditions for the simulated experiment). Next, we normalized each predictor to have unit variance (using the default setting of the fitting package, glmnet; Friedman et al., 2019). Last, we trained multiclass logistic regression on the normalized predictors (i.e. the differences in SS with added noise) with L2 penalization, and optimized the hyperparameter that weights the penalization by 10-fold cross-validation (i.e. the default in the glmnet package). For each experiment, we trained eight different models (observers), using different noise samples and different samples for the training set, leading to some variability between model observers.

After training the models, we tested their discrimination performance on a test set comprising 1500 texture pairs (500 of each class) for each surround condition.

We verified that all the trends and conclusions are robust to the choices of target size, penalization (we also tested elasticnet, which uses a mixed L1 and L2 penalization), and noise level. Furthermore, we also ran the model with a variation of the task that involved no stimulus sampling variability (see Supplementary section S6).

Statistical analysis

All experiments were first performed with texture T1, and all but experiment 1 were then reproduced with other textures. Experiments performed with T1 sometimes had more conditions than experiments with the other textures. These conditions exclusive to T1 are analyzed separately in the supplementary analysis.

We analyzed the data of the experiments and the simulations using generalized linear mixed models (GLMMs) of the binomial family (Gelman & Hill, 2006). In these models, we included a fixed effect for each parameter of interest and an offset term. For each of the fixed effects, we added random effects. When applying the GLMMs to multiple textures to estimate the mean effect across textures, we included for each fixed effect a random effect for texture and a random effect for participants nested within texture. We also applied the GLMMs to individual textures, both for the analysis of data that was only collected with one texture (e.g. experiment 1), and for estimating the effects of the different manipulations on each texture. In the plots showing the effects for multiple textures, the estimate for each individual texture was obtained by fitting a GLMM to that texture individually. In these cases, we only used a random effect for participants. Correlations between random effects in the model were always set to zero, to avoid overly complex models (Bates, Kliegl, Vasishth, & Baayen, 2015).

All the GLMMs fitted by maximum likelihood using the R package lme4 (Bates et al., 2019). The reported p value for each effect was obtained by a likelihood ratio test (LRT) between the full model and the null model, in which that fixed effect is set to zero. The 95% confidence intervals of the fixed effects were obtained by likelihood profiling.

The analysis in the text is based on the parameters fitted by these models, which are in log-odds ratio (LOR) units. Although less intuitive than simple differences between success probabilities, this is a more adequate measure for the experimental effects, especially given the variability in performance between participants and textures.

In some cases, we fitted a generalized linear model (GLM) to the data of each participant in an experiment in order to display the actual observed LOR for each individual (e.g. Figure 3). These models contained no random effects. The confidence intervals for the parameters were obtained by the Wald method, and their p values by the Wald test.

We excluded from analysis experimental sessions in which the participant performed below 45% correct for all conditions (chance level performance is 33%), to avoid strong floor effects. This criterion discarded 14 of the total 189 experimental sessions. In the main text, we report for each experiment the number of experimental sessions that satisfied the inclusion criterion. All results and analyses are robust to removing this exclusion criterion, as well as to excluding the main author from the analysis.

Data analysis was performed in R version 3.4.4 (R Core Team, 2018) using the packages lme4 1.1-19 (Bates et al., 2019), dplyr 0.7.6 (Wickham, François, Henry, & Müller, 2018), tidyr 0.8.1 (Wickham, Lionel Henry, & RStudio, 2018), ggplot2 3.0.0 (Wickham, 2016), broom 0.5.0 (Robinson & Alex Hayes, 2018), MASS 7.3-50 (Venables & Ripley, 2002), and knitr 1.20 (Xie, 2015).

Data availability

The anonymized raw data of the experiments, together with the analysis code, and the code for running the experiments, are available in the Open Science Framework (https://osf.io/8zr5h/). All participants gave informed written consent for their anonymized data to be publicly shared.

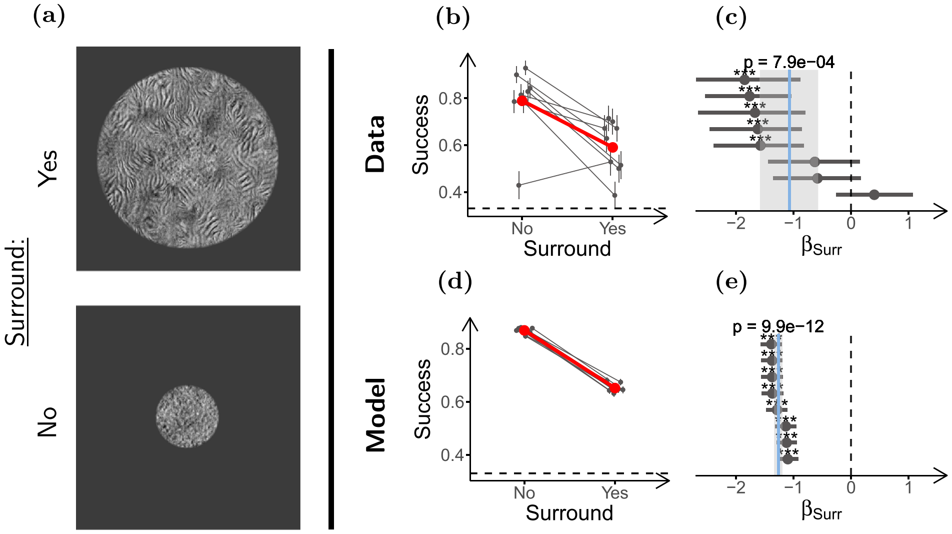


Figure 3. **Surround textures impair texture discrimination performance.** (**a**) Stimulus configurations used in the experiment (only scrambled targets shown). Top: Target with surround, and bottom: target without surround. (**b**) Task performances for the two conditions. The gray dots and lines show the performance of individual participants. Vertical lines indicate the ±SD of the estimated performance. Horizontal jitter was applied to aid visualization. The larger red dots show mean performance across participants for each condition. The dashed horizontal line shows chance performance. (**c**) Log odds ratios (LORs) between the presence and absence of the surround (β_{Surr}), estimated from the performance data in **b**. Each dot shows the LOR for one participant (estimated by fitting a GLMM), and the horizontal lines indicate their 95% confidence interval. Statistical significance of the LOR for the individual participants obtained by the Wald test is indicated as follows: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. The vertical solid blue line indicates the estimated mean LOR for the population (estimated by fitting a GLMM), and the grey shade indicates its 95% confidence interval. The p value of the mean LOR estimate as obtained by likelihood-ratio test (LRT) is indicated above the solid line. The dashed vertical line marks the LOR at which there is no difference between the conditions. (**d**, **e**) show the same as **b** and **c** but for the model observers. Participants ($n = 8$) performed 70 trials in each condition, and model observers ($n = 8$) discriminated 1500 stimuli per condition.

Results

We used a PS texture discrimination task (details in Figure 2 and Methods) to study contextual modulation of texture perception in peripheral vision. We refer to contextual modulation as the observed phenomenon by which perception of a part of a visual stimulus is affected by its surrounds, regardless of the precise underlying mechanisms. In our experiments, we measure changes in contextual modulation as the changes in task performance between conditions with different surrounds (taken as indicative of changes in target perception between conditions induced by these surrounds). PS textures are characterized by a set of SS inspired in natural image statistics and early human vision, including the correlations between the outputs of V1-like filters selective for orientation and spatial frequency. The corresponding SS model implementation consists of two stages: the first stage computes the responses of the V1-like filters to the input image, and the second stage evaluates the PS statistics of those filter activations within fixed pooling windows (see Figure 1).

The task required discriminating patches of naturalistic PS texture from their corresponding phase scrambled textures (see Figure 2) in a three alternative forced choice design (we refer to the patches to be discriminated as targets). These PS and phase-scrambled texture pairs have the same FAS, which means they activate the V1-like filters of the SS model with the same average energy, and are thus matched in the first stage of the SS model. Unlike phase-scrambled textures, PS textures also have a more structured distribution of filters activations, corresponding to HOS that drive the second stage of the SS model and lead to a more natural appearance (Portilla & Simoncelli, 2000).

To evaluate whether our experimental observations could be captured by the feedforward SS model with fixed pooling windows, we implemented a model SS observer to solve the task using a linear classifier on the PS statistics of the stimuli, computed over a fixed area centered on the target (see Figure 2b, Methods). We then compared qualitatively the model's discrimination performance to the participants. The radius of the pooling windows was chosen according to Bouma's law of crowding, which says that surrounding stimuli can interfere with target identification when they are within a distance of approximately 0.5 times the target eccentricity (Pelli, Palomares, & Majaj, 2004; see Methods section).

The results are divided into three sections. First, we report the effect of the surround on performance, and its dependence on target-surround grouping or segmentation. Then, we explore the relevance of the statistical structure of the surround texture to contextual modulation. Last, we study the relation of this contextual interaction to crowding.

Contextual modulation and grouping

Target-surround grouping, or conversely segmentation, is a major modulator of contextual inter actions in vision, especially for crowding (Levi, 2008; Saarela & Herzog, 2009; Manassi et al., 2013; Qiu et al., 2013). It has been argued that these segmentation and grouping processes are an important missing component in pooling models of peripheral vision, including the SS model (Manassi et al., 2013; Doerig et al., 2019; Wallis et al., 2019). Despite considerable work using stimuli, such as objects, shapes, or features (e.g. Kooi et al., 1994; Saarela, Westheimer, and Herzog, 2010; Manassi et al., 2013; Manassi et al., 2016), our understanding of how grouping processes affect peripheral perception is still incomplete because it is not clear how to relate those tasks that use non-texture stimuli to the SS model, which may be affected by more global stimulus information (Rosenholtz et al., 2019), and whether those results extend to texture processing.

Thus, to better understand the role of grouping and segmentation in the SS model, and how they influence perception of textures, we sought to determine whether contextual modulation of naturalistic texture perception is affected by segmentation or grouping cues.

Experiment 1: Target-surround discontinuity reduces contextual modulation

First, we measured whether naturalistic texture perception is affected by contextual modulation. Based on the relevance of contextual modulation for target identification in peripheral vision, we expected task performance to be impaired by surrounding textures. To test this, we presented participants ($n = 8$) with targets in isolation, and with targets surrounded by an uninformative texture ring that was sampled from the same PS texture (see Figure 3a).

As expected, task performance was considerably worse for the surrounded targets (see Figure 3b). To quantify the effect sizes and test for their statistical significance, we fitted a GLMM to the data (which allows to take into account between-participant variability; see Methods and Supplementary section S5). We report the LOR between the conditions (denoted by β), which is a measure of their difference in success probability (see a guide for converting between the two in Supplementary section S5). For example, β_{Surr} quantifies the effect of the surround around the target, and $\beta_{Surr} < 0$ means that the surround hindered performance. Figure 3c shows that, in our experiments, the surround strongly impaired performance, and that the effect was statistically significant ($\beta_{Surr} = -1.07$, ci $= [-1.58$ to $-0.57]$, $p = 8 \times 10^{-4}$). This effect was captured by our implementation of the SS model (see Figure 3e).

We next tested whether segmentation affects this contextual modulation, and whether the effect can be captured by our SS model implementation. To probe the effect of segmentation, we presented participants ($n = 9$) with two kinds of stimuli, either with continuous target and surround, or with a visible gap that induced target surround segmentation (Figure 4a). Importantly, the gap was generated by shrinking the target of the continuous stimuli, keeping surround geometry the same in the two conditions. With this design, if pooling regions are constant, the two conditions would have the same amount of surround texture pooled with the target, but in the discontinuous condition there would be less target texture to be integrated (due to the smaller target size). In line with what could be expected from the ratio of informative target texture and uninformative surround texture for each stimulus, our implementation of the SS model showed worse performance in the discontinuous than in the continuous condition (Figures 4d, 4e; a similar reasoning to that applied in Manassi et al., 2013). This is in contrast to what we expect from previous studies using simple stimuli, in which segmentation reduced contextual modulation (Kooi et al., 1994; Saarela et al., 2010; Manassi et al., 2012; Manassi et al., 2013; Qiu et al., 2013; Manassi et al., 2015; Manassi et al., 2016). Figure 4b shows that performance increased moderately when target and surround were discontinuous ($\beta_{Discont} = 0.62$, ci $=$ $[0.37$ to $0.87]$, $p = 1 \times 10^{-4}$; see Figure 4c). Thus, our SS model implementation was unable to capture the effect of segmentation (see Supplementary section S4 for further discussion).

We also found that the observed effect of discontinuity is sensitive to the size of the gap (see Supplementary section S2), likely because the gap size affects gap visibility, and also the difference in target sizes between the conditions. In addition, notice that segmentation did not completely remove contextual modulation, that is, performance was still lower for the discontinuous surround than for the target alone ($\beta^{Discont}{}_{Surr} = -0.40$, ci $= [-0.68$ to $-0.15]$, $p = 5 \times 10^{-3}$; Supplementary section S2).

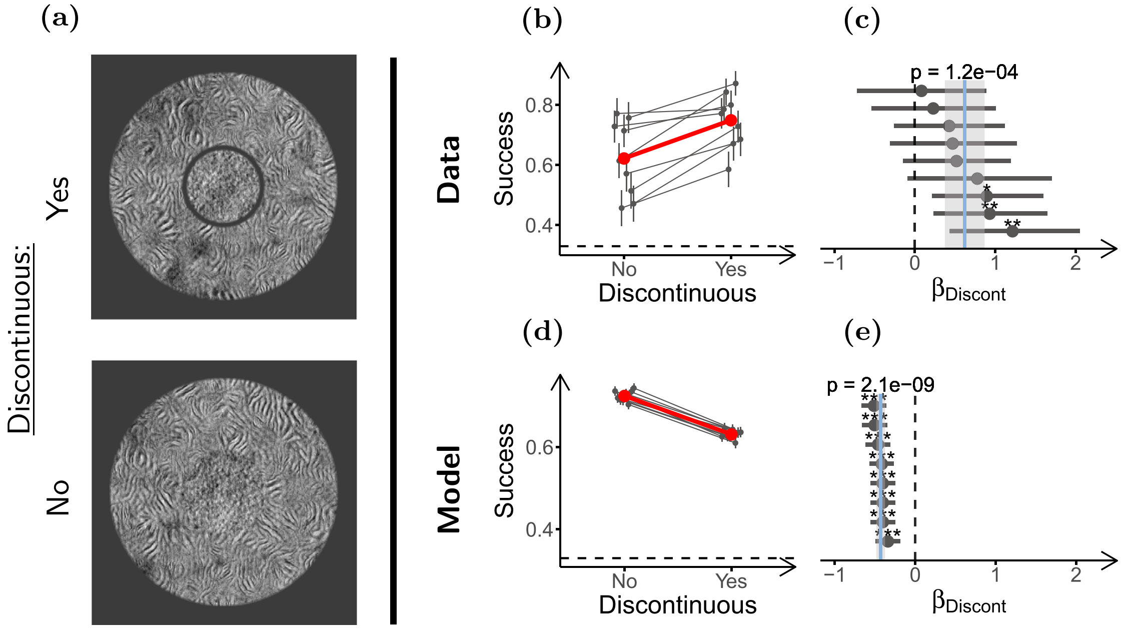


Figure 4. **Segmentation reduces contextual modulation.** (**a**) Stimulus configurations used in the experiment (only scrambled targets shown). Top: Discontinuous stimulus (smaller target size), and bottom: Continuous stimulus (larger target size). (**b**) Task performance for the two conditions. (**c**) LOR for discontinuity ($\beta_{Discont}$), estimated from the performance data in **b**. (**d**, **e**) Same as **b** and **c** but for the simulated observers. Participants ($n = 9$) performed 70 trials in each condition, and model observers ($n = 8$) discriminated 1500 stimuli per condition. Panels **b** through **e** use the same conventions as Figure 3.

Experiment 2: The effect of target-surround discontinuity is mediated by segmentation

We reasoned that the gap between target and surround used to induce segmentation may also affect performance by other mechanisms, such as reducing the uncertainty of target location within the stimulus, or altering the SS of the stimulus in a way that is not captured by our SS model implementation. For example, it has been proposed that some uncrowding results may be explained by a better encoding or decoding of target information from the SS of the stimuli, allowed by the specific stimulus configurations that generate uncrowding (Rosenholtz et al., 2019). Given that the gap in our stimuli is colocalized with the target, it is possible that their low-level features induce changes in the SS that allow for a better decoding of target information (see Supplementary section S6 for modeling results suggesting that such factors may be relevant).

To control for the possible cues related to the gap but not to segmentation, we introduced a different target shape (split-target) consisting of two adjacent semicircles with their straight sides facing outward (see Figure 5a). This split-target shape had approximately the same texture area as the original disk target, and a gap could be introduced around its curved sides, while preserving target-surround continuity on the straight sides of the target.

Although the circular targets and the split targets had gaps with similar low-level properties, we expected no segmentation for the split-target stimulus because target-surround continuity is maintained. Thus, if the effect observed in the previous section was mediated by segmentation, we should find lower performance for the grouped continuous stimulus (split-target) as compared to the segmented discontinuous stimulus (disk-target). If the effects were mostly due to other factors introduced by the low-level properties of the gap, then we would expect similar performance for these two kinds of stimuli.

We presented participants ($n = 25$) with the disk-target and split-target stimuli using five different textures to verify that the results did not depend on a specific texture (most participants were shown only some of the textures, see Methods). Participants completed 40 experimental sessions (an experimental session consists of a participant completing the experiment with one texture) that satisfied the inclusion criterion (see Methods). Consistent with a role of segmentation in contextual modulation of texture perception, performance was moderately worse for the continuous (split-target) than for the discontinuous stimulus ($\beta_{Discont} = 0.42$, ci = [0.23 to 0.62], $p = 2 \times 10^{-3}$; see Figure 5c). In contrast to this observation, our implementation of the SS model showed little difference between the stimuli, showing again a failure to capture the segmentation effect (see Figure 5e).

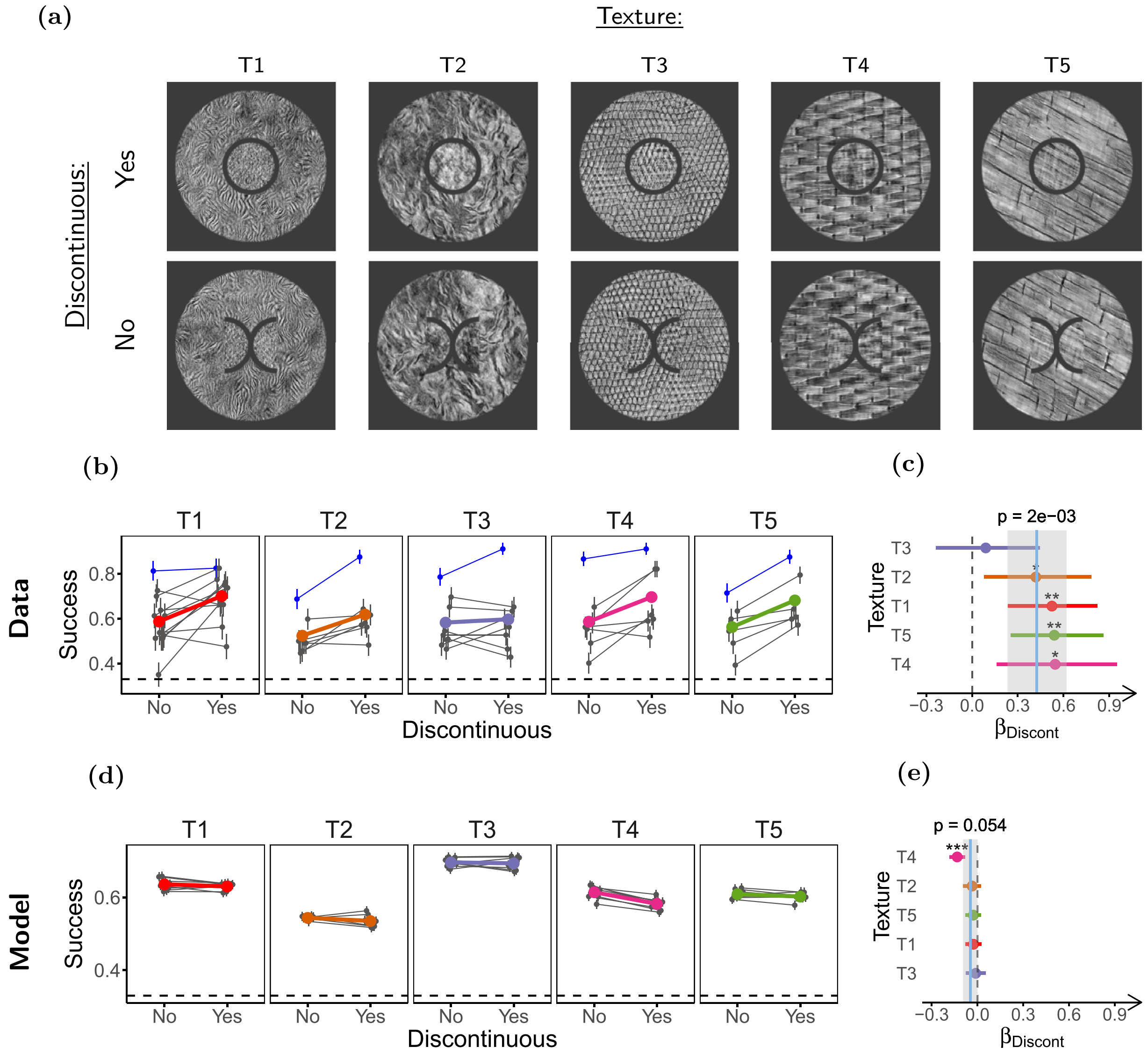


Figure 5. **Low level properties of the gap do not explain the effect of discontinuity.** (**a**) Stimuli used in the experiment. Top: Disk targets (discontinuous), and bottom: split-targets (continuous). (**b**) Task performance. Each panel shows the results for a different texture, with texture identity indicated above the panel. The layout of each panel is the same as in 3b, except a different color is used to identify the mean performance for each texture. The data of author DH are indicated by the blue symbols. (**c**) LOR for target-surround discontinuity ($\beta_{Discont}$). The colored dots show the LOR obtained by fitting a GLMM for each individual texture (color coded as in **b**) and the horizontal lines indicate their 95% confidence interval. The p value for the ($\beta_{Discont}$) of each individual texture, estimated by LRT, is indicated as follows: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The vertical solid blue line shows the mean $\beta_{Discont}$ across textures and participants estimated by a hierarchical GLMM model using all textures, and the shaded gray region shows its 95% confidence interval. The p value for this estimate obtained using LRT is indicated above the line. The dashed vertical line marks the value at which there is no difference between conditions. (**d**, **e**) Same as **b** and **c** but for the model observers. Participants ($n = 25$) completed 40 experimental sessions (see Methods), and performed between 80 and 112 trials per condition. Model observers ($n = 8$) discriminated 1500 trials per condition.

We also verified, using additional stimuli for texture T1 (see Supplementary section S6), that splitting the target had a small and nonsignificant effect on performance ($\beta_{Split} = -0.09$ ci $= [-0.31$ to $0.14]$, $p = 0.43$ see Supplementary section S6) validating the use of this experiment to control for low-level gap properties. Furthermore, the estimated effect of the gap after accounting for segmentation was also close to 0 ($\beta_{Gap} = 0.04$, ci $= [-0.15$ to $0.22]$, $p = 0.71$, see Supplementary section S6), suggesting that effects of the gap other than inducing target-surround segmentation are negligible in our task. This extended analysis supports the interpretation that the effect of segmentation on contextual modulation cannot be wholly explained by the changes in target encoding or decoding allowed by its colocalization with the gap, although it remains possible that a more complex SS model with more statistics or more complex structure could capture these results.

We conclude from these experiments that target-surround segmentation is an important factor in mediating contextual modulation of texture perception, that a discontinuity between target and surround induces segmentation and thus reduces contextual modulation, and that this effect is not observed in our implementation of the feedforward SS model with fixed pooling windows.

Effect of surround statistics

Besides the geometric cue (the gap) we considered above, another important factor that can reduce contextual modulation is target-surround dissimilarity. The effect of target surround dissimilarity is well reported for object and feature crowding, where the effects of the surround on target identification can be reduced if the two differ in aspects such as color, orientation, or higher-level attributes (Kooi et al., 1994; Louie, Bressler, & Whitney, 2007; Põder, 2007; Farzin, Rivera, & Whitney, 2009; Whitney & Levi, 2011; Manassi & Whitney, 2018), thus increasing target saliency (Gheri, Morgan, & Solomon, 2007). This breakdown in statistical similarity is known to enhance perceptual saliency (Li, 1999; Li, 2002) and in some cases is suggested to act through segmentation (Whitney & Levi, 2011; Manassi et al., 2013). Understanding the effects of surround structure on contextual modulation of texture perception is important because during natural scene perception there is abundant variability in texture properties and arrangement. Furthermore, different levels of surround structure are often used as proxies for different stages of neural processing (Louie et al., 2007; Farzin et al., 2009; Gong, Xuan, Smart, & Olzak, 2018; Manassi & Whitney, 2018), which could provide insights on the mechanisms behind our observations. For these reasons, we next asked how target-surround dissimilarity affects peripheral texture perception, and how it interacts with segmentation.

Experiment 3: FAS dissimilarity but not hos dissimilarity strongly reduces contextual modulation through segmentation

We focused on target-surround dissimilarity at the FAS and HOS levels because they are related to the SS model (Freeman & Simoncelli, 2011; Freeman et al., 2013) and to physiology (Balas et al., 2009; Freeman & Simoncelli, 2011; Freeman et al., 2013; Okazawa et al., 2015; Ziemba et al., 2016; Okazawa et al., 2017), as discussed above. Previous work on contextual modulation (Xing & Heeger, 2000; Whitney & Levi, 2011; Manassi & Whitney, 2018) suggests that dissimilar surrounds should have a smaller influence on target perception. In addition, it is well known that textures can be segmented from one another based on dissimilarity in their statistics (Julesz, 1962; Rosenholtz, 2014; Victor, Conte, & Chubb, 2017), and thus we expect that textures that allow for good target-surround segmentation will lead to reduced contextual modulation. However, although effects of FAS and certain HOS in perceptual segmentation and contextual modulation have been studied in a variety of experimental settings (e.g. Julesz, Gilbert, & Victor, 1978; Julesz, 1962; Julesz & Caelli, 1979; Xing & Heeger, 2000; Whitney & Levi, 2011; Victor et al., 2013; Hermundstad et al., 2014; Zavitz & Baker, 2014; Victor et al., 2017), and a wide arrange of computational models attempt to explain texture segmentation and contextual modulation (for reviews and examples see Bergen and Landy, 1991; Li, 2002; Thielscher & Neumann, 2005; Bhatt, Carpenter, & Grossberg, 2007; Thielscher, Kölle, Neumann, Spitzer, & Grön, 2008; Landy, 2013; Rosenholtz, 2014; Victor et al., 2017), these processes have not been systematically studied for naturalistic textures, and their effects can also be task dependent (Vancleef et al., 2013; Victor et al., 2017), making it difficult to tell a priori what effects they may have in our task.

To test the effects of FAS and HOS dissimilarity, we compared three different surround textures (Figure 6a): (1) the same PS texture as the target (none dissimilar), (2) a different PS texture with FAS and pixel histogram matched to the target PS texture (HOS dissimilar), and (3) a different PS texture with only its pixel histogram matched to the target (FAS and HOS dissimilar). Furthermore, to study the interaction of FAS and HOS dissimilarity with segmentation, we showed these surround textures in both the continuous and discontinuous conditions. In this experiment, target size was the same for the continuous and discontinuous conditions, and the gap was generated by enlarging the surround for the discontinuous condition (increasing inner and outer diameter to maintain its width).

We presented participants ($n = 22$) with 4 different target textures (see Figure 6a), adding to 31 experimental sessions. To analyze the data, we fitted a GLMM with parameters for FAS dissimilarity (β_{FAS}) and HOS dissimilarity (β_{HOS}) separately to the continuous and discontinuous conditions, where the effect of FAS dissimilarity is estimated as the change in performance between the condition of HOS dissimilarity and the condition of HOS and FAS dissimilarity.

First, we asked whether the two levels of dissimilarity had an effect for the continuous stimulus. The effect of

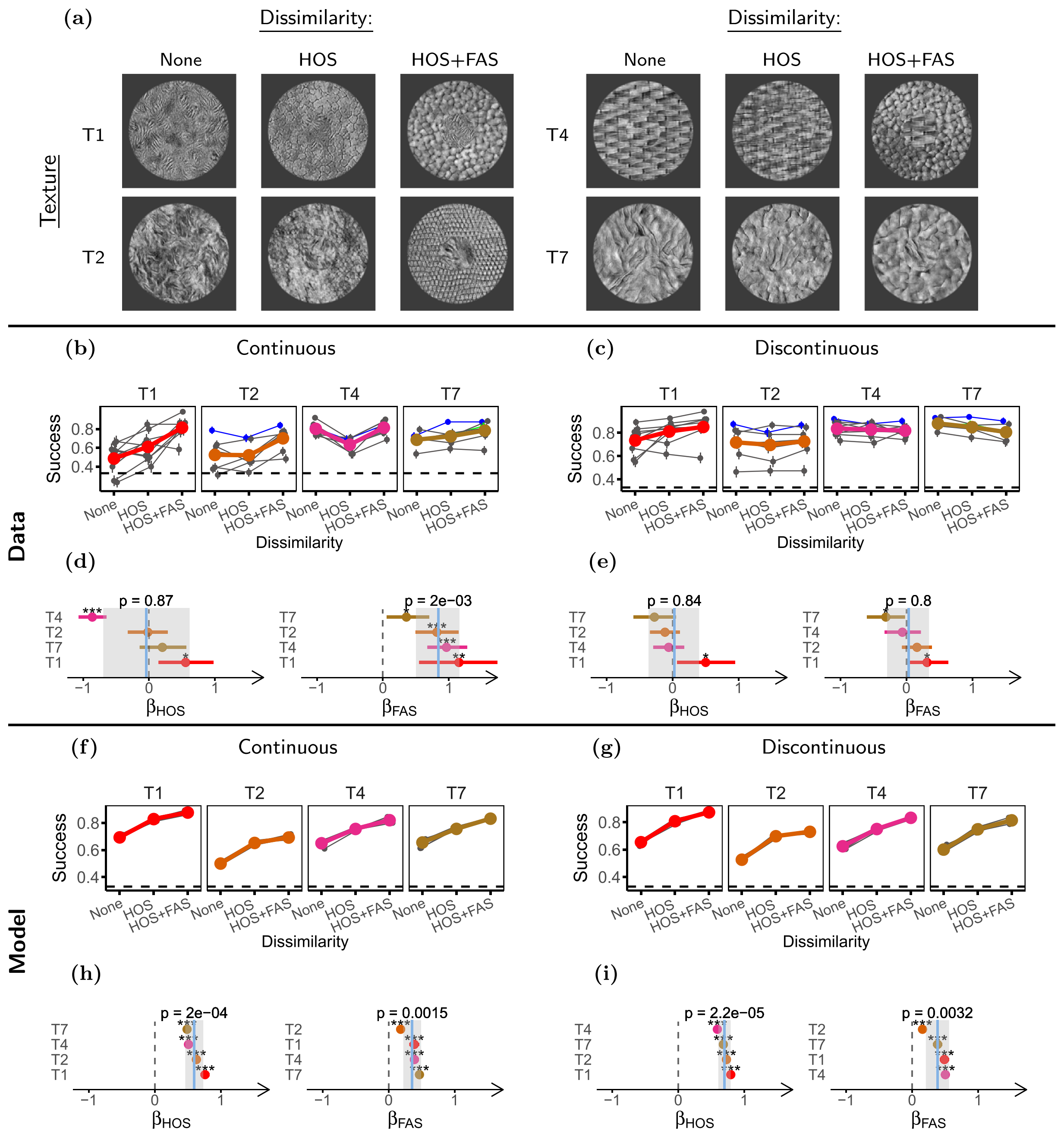


Figure 6. **Target-surround dissimilarity reduces contextual modulation.** (**a**) Samples of the stimuli used in this experiment, showing for target textures the three levels of target-surround dissimilarity used in the experiment (discontinuous stimuli not shown). (**b**, **c**) Task performances for the different target-surround dissimilarities in the continuous and discontinuous conditions, respectively. (**d**, **e**) LOR for HOS (β_{HOS}) and FAS (β_{FAS}) dissimilarity in the continuous and discontinuous conditions respectively. (**f–i**) Same as **b** through **e** but for the model observers. Participants ($n = 22$) completed 31 experimental sessions, and performed between 60 and 120 trials per condition. Model observers ($n = 8$) discriminated 1500 stimuli per condition. The plots in this figure use the same conventions as the corresponding plots in Figure 5.

HOS dissimilarity was close to 0 and not significant ($\beta^{Cont}{}_{HOS} = -0.04$, ci = [−0.69 to 0.61], $p = 0.87$; see Figure 6d), whereas FAS dissimilarity generated strong improvements in performance overall ($\beta^{Cont}{}_{FAS} = 0.84$, ci = [0.50 to 1.16], $p = 2 \times 10^{-3}$; see Figure 6d).

We note that the effect of HOS showed considerable variability between textures. In particular, for texture T4 performance was strongly reduced for dissimilar HOS, contrary to expectations. This is likely because the surround without dissimilarity for this texture has a high regularity that introduces a phase effect at the target-surround boundary, which could act as a segmentation cue.

To better understand the relation between dissimilarity and segmentation, we then asked whether dissimilarity interacted with discontinuity. If the effects of dissimilarity are mediated simply by surround statistics pooled over fixed regions, we would expect dissimilarity effects for the discontinuous condition comparable to those of the continuous condition (assuming, as we do, a pooling area with the radius of Bouma's law such that the small change in surround geometry is negligible). On the other hand, if dissimilarity effects are mediated by segmentation, we expect the effects to be reduced in the discontinuous condition where segmentation is already induced by the gap. Consistent with the second mechanism, we found that target-surround dissimilarity had little effect on contextual modulation in the discontinuous condition (see Figure 6c) for both HOS (${\beta^{Discont}}_{HOS} = 0.03$, ci = [−0.36 to 0.40], $p = 0.84$; see Figure 6e), and FAS (${\beta^{Discont}}_{FAS} = 0.03$, ci = [−0.29 to 0.34], $p = 0.80$; see Figure 6e), although there was considerable variability between textures. We verified that the change of the effect of FAS dissimilarity for the discontinuous condition was significant (see Supplementary section S4 and Supplementary Figure S6).

Our analysis therefore suggests that FAS dissimilarity effects are strong and mediated by segmentation, whereas HOS dissimilarity effects show considerable variability across textures but are, on average, weak. We then tested whether these results could be captured by our implementation of the SS model. First, in the continuous condition our model showed a strong improvement in performance when there was HOS dissimilarity, and much weaker changes for FAS dissimilarity (see Figures 6f, 6h). Second, these effects were mostly unchanged for the discontinuous condition (see Supplementary section S6), due to the lack of explicit segmentation processes. Therefore, our implementation of the SS model was not able to capture the patterns observed in the human data.

Experiment 4: Naturalistic structure in the surround is important to recruit contextual modulation

The results of the previous section show that a surround with different naturalistic HOS than the target can still exert substantial contextual modulation. Interestingly, other studies have previously shown that contextual modulation can be reduced by removing the natural HOS from the surround. Perceptually, this has been observed for tasks involving recognition and discrimination of natural scenes in peripheral vision (Wallis, Bethge, & Wichmann, 2016; Gong et al., 2018), and for local orientation processing during scene perception (Neri, 2017). Neurally, it has been shown that phase-scrambling the surround (i.e. the HOS are removed but the FAS maintained) strongly affects contextual modulation of neural activity in response to natural images in V1 (Guo, Robertson, Mahmoodi, & Young, 2005; Pecka, Han, Sader, & Mrsic-Flogel, 2014; Coen-Cagli, Kohn, & Schwartz, 2015) and to naturalistic textures in V2 (Ziemba, Freeman, Simoncelli, & Movshon, 2018). This effect of naturalness is thought to reflect that contextual modulation is tuned to natural image statistics, to support efficient coding and optimal perceptual inferences (Pecka et al., 2014; Coen-Cagli et al., 2015). This interpretation seems also in line with previous work with artificial textures, proposing that the asymmetries between textures with uniform and random orientation in texture filling-in could be related to a process of perceptual inference (Hindi Attar, Hamburger, Rosenholtz, Götzl, & Spillmann, 2007). In other work using natural and phase-scrambled scenes, the effect of phase-scrambling has been explained (Gong et al., 2018) as resulting from a weaker engagement of higher areas in the visual hierarchy, leading to reduced contextual modulation in these higher areas. In the context of this literature, our finding of a relatively weak effect of HOS dissimilarity in the previous experiment raises the question of whether the presence of natural HOS is necessary for recruiting contextual modulation for textures.

To address this question, we compared the effects of naturalistic and phase-scrambled surrounds continuous to the target (Figure 7a). Because our experiments required to identify the phase-scrambled target, we reasoned that target-surround similarity with the texture to be identified could affect contextual modulation and lead to unpredictable confounding effects. Therefore, to balance out this possible effect of similarity, we asked half the participants to identify the phase-scrambled texture and the other half to identify the naturalistic texture (modifying the task accordingly, see Methods), and we report the results from both task variants together.

Participants ($n = 28$) were presented with 5 textures, adding to 43 experimental sessions. Consistent with previous studies, we observed that performance was worse with natural HOS in the surround ($\beta_{Nat} = -0.91$, ci = [−1.25 to −0.56], $p = 10 \times 10^{-4}$; see Figure 7c). This is in agreement with previous physiology studies (Guo et al., 2005; Pecka et al., 2014; Coen-Cagli et al., 2015) showing that naturalistic HOS in the surround are important for fully engaging contextual modulation, possibly due to the tuning of contextual modulation to natural image statistics for efficient coding and inference. Together, these results and those from experiment 3 suggests that although the presence of HOS in the surround is important for contextual modulation, their similarity to the HOS of the center is of secondary importance. We note, however, that

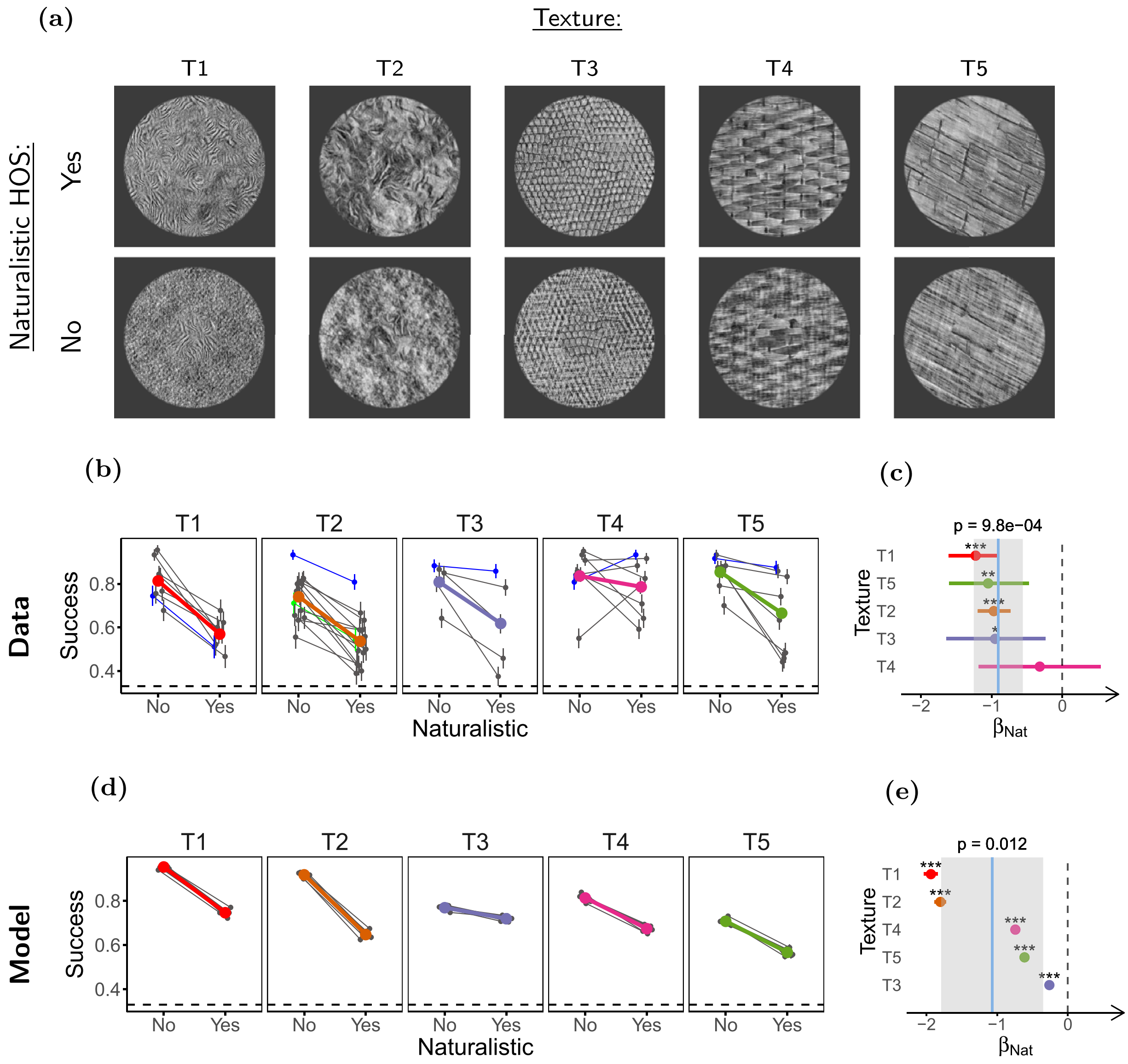


Figure 7. **Naturalistic HOS increase contextual modulation.** (**a**) Stimuli used in the experiment. Top row: Naturalistic surrounds. Bottom row: phase-scrambled surrounds (only naturalistic targets are shown). (**b**) Task performance. (**c**) LOR for the presence of naturalistic HOS (β_{Nat}). (**d**, **e**) Same as **b** and **c** but for the model observers. Participants ($n = 43$) completed 43 experimental sessions and performed between 90 and 120 trials per condition. Model observers ($n = 8$) discriminated 1500 trials per condition. The plots in this figure use the same conventions as the corresponding plots in Figure 5.

contextual modulation still occurs for phase-scrambled surrounds (Supplementary section S7), thus the phenomenon can occur in the absence of naturalistic HOS.

Although, as discussed, the effect of naturalness may reflect the tuning of contextual modulation to natural statistics (Pecka et al., 2014; Coen-Cagli et al., 2015; Ziemba et al., 2018), we observed a qualitatively similar effect of naturalness in our SS model implementation (see Figure 7e). This means that at least part of the effect of naturalness could be mediated by simple pooling. Nonetheless, for textures T1 and T2, we also studied the interaction between naturalness and segmentation (see Supplementary section S5, Supplementary Figure S7), and found that adding a discontinuity reduced the effect of naturalness ($\beta_{Nat:Discont} = 0.43$, ci = [0.06 to 0.83], $p = 0.04$; see Supplementary section S7), whereas this effect was not captured by our SS model ($\beta_{Nat:Discont} = 0.05$, ci = [−0.03 to 0.12], $p = 0.15$; Supplementary section S7). In addition, further analysis of the model shows that the observed naturalness effect in the model is not due to surround naturalness itself, but rather due to some specific features of our stimulus generating process (see Supplementary section S6, Supplementary

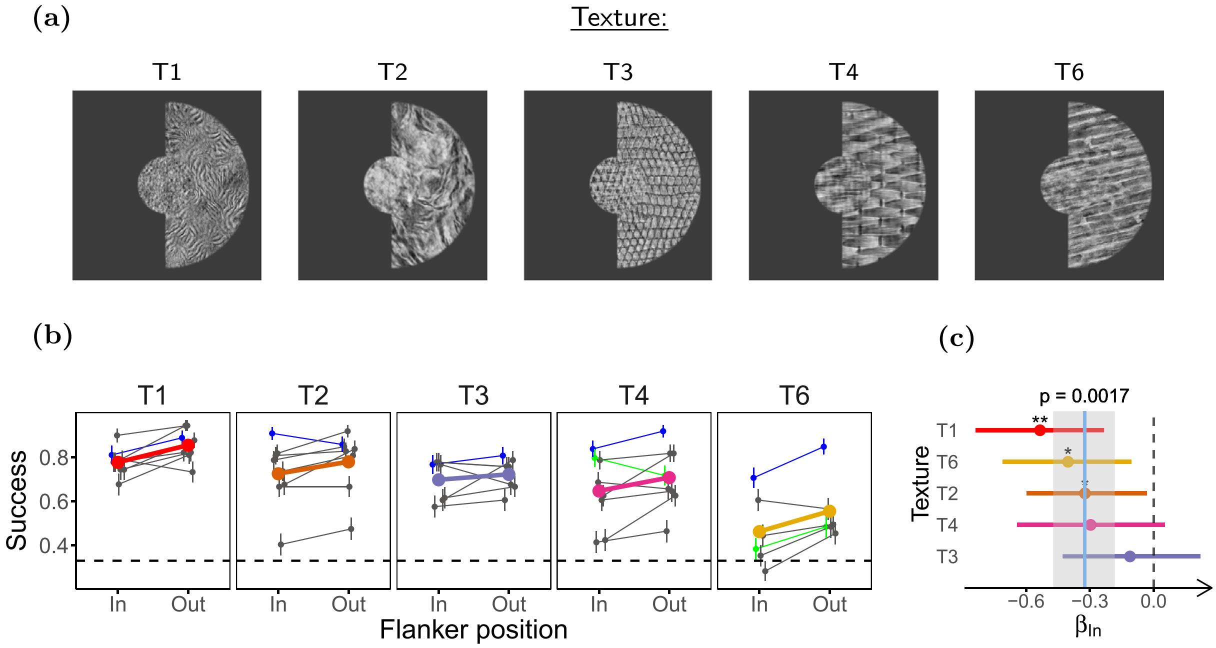


Figure 8. **Inward surrounds affect performance more than outward surrounds.** (**a**) Stimuli used in the experiment. Inward and outward surround conditions differ in the position of the half ring of surround texture relative to the fixation point. (**b**) Task performance for the different surround positions. (**c**) LOR for inward versus outward surround (β_{In}). Participants ($n = 21$) completed 40 experimental sessions, and performed between 90 and 99 trials per condition. The figure uses the same conventions as Figure 5.

Figure S12). Thus, it is likely that pooling is not the only mediator of the effect of naturalness in our experiments.

In conclusion, these results suggest that naturalistic HOS are important for fully engaging contextual modulation phenomena. This is compatible with suggestions that neuronal contextual modulation phenomena are tuned to the structure of natural images (Pecka et al., 2014; Coen-Cagli et al., 2015), and more specifically, with the results observed for neuronal contextual modulation phenomena in V and V2, that may be mechanistically related to our results (see Discussion).

Texture crowding

We have thus far shown that texture perception is affected by contextual modulation, and influenced by segmentation and target surround dissimilarity. These characteristics are consistent with a possible role of visual crowding, a contextual modulation phenomenon often regarded as the most important factor of peripheral vision (Rosenholtz, 2016). The SS model explains crowding as a loss of information from pooling together target and surround features when computing local SS (Balas et al., 2009; Freeman & Simoncelli, 2011; Whitney & Levi, 2011; Freeman et al., 2013). However, it is not clear whether this explanation, that is often applied to tasks on non-texture stimuli, should hold for our task. Thus, we decided to test whether the contextual modulation we observed is due to crowding.

There are two main diagnostic criteria for crowding. One is compliance with Bouma's law, which states that the critical distance at which surrounds interfere with target perception scales linearly with eccentricity with a slope of approximately 0.5 (Pelli et al., 2004). The other is an inward-outward asymmetry in which surrounds more eccentric (outward) to the target exert a stronger modulation than surrounds more central (inward) to the target (Pelli et al., 2004; Petrov, Popple, & McKee, 2007; Farzin et al., 2009; Whitney & Levi, 2011; Rosenholtz, 2016). Probing Bouma's law with textures poses experimental challenges, such as changing target-surround distance without breaking continuity or altering target size, and determining how to measure distance between texture stimuli (e.g. Rosen, Chakravarthi, & Pelli, 2014). Therefore, we decided to probe the characteristic inward-outward asymmetry of crowding.

Experiment 5: Effect of surround position is small, highly variable, and task dependent

To test for inward-outward asymmetry in our task, we used half-ring-shaped surrounds (Figure 8a) placed inward or outward of the target. Participants ($n = 21$) were presented with 5 different textures, completing 37 experimental sessions. Opposite to

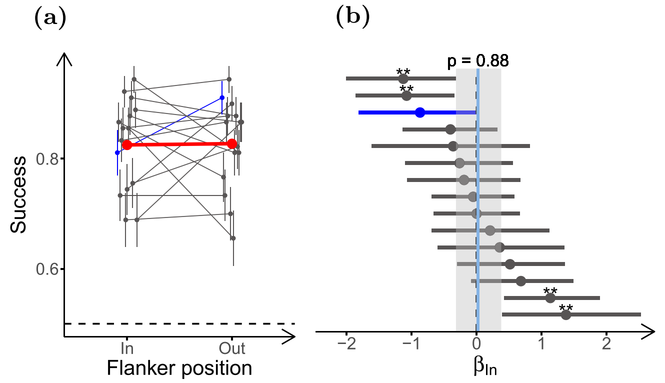


Figure 9. **Reduced inward-outward asymmetry with single target.** (**a**) Task performance for the different surround positions for the task using only one target, for texture T1. (**b**) LOR of the position of the surround for the task using only one target (β_{In}). Participants ($n = 15$) performed 90 trials per condition. The plots in this figure use the same conventions as the corresponding plots in Figure 3.

what has been reported in most crowding studies, performance in our task was consistently lower when the surround was inward of the target ($\beta_{In} = -0.32$, ci = [−0.47 to −0.18], $p = 2 \times 10^{-3}$; see Figure 8c). This suggests that crowding as reported for classical letter detection or orientation discrimination may not be the main contextual modulation phenomenon in our experiments.

Nonetheless, unlike the task used here, most reports of inward-outward asymmetry use only one target (Banks, Larson, & Prinzmetal, 1979; Petrov et al., 2007; Farzin et al., 2009; Manassi et al., 2012). To verify that the previous result is not only due to this task-related effect, we repeated the experiment for texture T1 using only one target, presented to the right of the fixation point. Participants ($n = 15$) had to report whether the target was naturalistic or phase scrambled. Using this new task, we observed an effect of surround position close to 0 ($\beta_{In} = 0.02$, ci = [−0.32 to 0.38], $p = 0.88$; Figure 9b). We also verified whether this lack of an effect is due to easier task conditions that bring performance to ceiling levels by using an unsurrounded control condition. Performance was significantly lower for the surrounded than for the control condition in this experiment ($\beta_{Surr} = -0.46$, ci = [−0.72 to −0.20], $p = 1 \times 10^{-3}$), meaning that the lack of an effect was not due to ceiling performance. This lack of inward-outward asymmetry is not what would be expected from the classical asymmetry in crowding, and thus supports the conclusion from the experiment using two targets. Nonetheless, we also note that the difference between the results from the two tasks is in agreement with an effect of task and attention on inwards-outwards asymmetry, such as shown in a previous study in which biasing attention toward the center of the visual field inverted the direction of inward-outward asymmetry (Petrov & Meleshkevich, 2011b).

Despite the lack of a clear asymmetry in the average performance, variation between participants was high, and some individual participants showed strong effects of surround position in both directions. One plausible interpretation of this result is that contextual modulation in our task arises from different contributing processes (e.g. crowding and surround suppression, although others processes are possible; see Discussion) and that participants with stronger crowding effects would show worse performance for outward targets, whereas participants more affected by other processes would show little or opposite asymmetry. This hypothesis is in line with previous studies reporting substantial variability in sensitivity to crowding between observers (Kooi et al., 1994; Petrov & Meleshkevich, 2011a; Wallace, Chiu, Nandy, & Tjan, 2013; Lev & Polat, 2015). In addition, we hypothesize that this variability in sensitivity to contextual modulation phenomena could arise from the use of different strategies for solving the task, possibly contributing to the considerable between-participant variability that we observed in the results of the previous experiments.

In conclusion, these results suggest that the processes that underlie crowding in experimental paradigms, such as letter recognition, and that have been widely reported to be stronger for outward surrounds, interact with other processes of at least comparable relevance to contextual modulation of texture perception, that show little or the opposite inward-outward asymmetry.

Discussion

Although the SS model of peripheral vision has had considerable success (Rosenholtz, 2016), studies using complex scenes (Wallis et al., 2019) and simple object-like stimuli (Saarela et al., 2009; Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Francis et al., 2017; Doerig et al., 2019) suggest that including processes of segmentation and grouping together with contextual modulation is crucial for a more accurate understanding of peripheral vision. Here, we showed that PS texture perception in the periphery is modulated by spatial context, and that contextual modulation is strongly reduced by segmentation engaged both by a gap between target and surround, and by target surround dissimilarity (see Figures 4, 5, 6). Although the relevance of segmentation and target-surround dissimilarity for contextual modulation has been studied for discrimination tasks using simple features or objects (Kooi et al., 1994; Zenger-Landolt & Koch, 2001; Sayim, Westheimer,

& Herzog, 2008; Saarela & Herzog, 2009; Saarela et al., 2010; Whitney & Levi, 2011; Manassi et al., 2013; Qiu et al., 2013; Manassi & Whitney, 2018), this is, to our knowledge, the first report of such effects for texture discrimination, which likely involves different processing of the visual input (Cant, Large, McCall, & Goodale, 2008; Cavina-Pratesi, Kentridge, Heywood, & Milner, 2010; Cant & Xu, 2012; Rosenholtz, 2014). Furthermore, although the simple feature and object stimuli are more difficult to relate to the SS model (Rosenholtz et al., 2019), our choice of stimuli and task allowed for a direct comparison with the SS model.

In line with previous work using a vernier discrimination task to show that adding more flankers could reduce crowding if these favored target segmentation (Malania, Herzog, & Westheimer, 2007; Manassi et al., 2012; Manassi et al., 2013), we found that increasing target size in our texture task can reduce performance if it eliminates a segmentation cue, and that this was not explained by our implementation of the SS model (see Figure 4). In addition, in line with similar work showing that the precise configuration of the surround is important because it determines grouping with the target (Manassi et al., 2013; Manassi et al., 2016), we show that the precise configuration of the target is important for the same reason. Our SS model implementation was not able to account for this effect (see Figure 5). These results thus support the view that the two-stage model with filtering followed by fixed pooling windows cannot fully explain crowding. We note, however, that this does not argue against the importance of SS as a general framework for understanding peripheral vision, but rather for the need to incorporate segmentation and flexible pooling processes more explicitly. As has been pointed out for previous studies (Rosenholtz et al., 2019), it is possible that a more sophisticated feedforward SS model (e.g. with a nonlinear decoder) could account for some of our segmentation results, leveraging the segmentation cues to extract relevant information from the SS of the stimulus. To test for this, we introduced in experiment 2 and in Supplementary section S6 a control for some of the major ways in which this could happen, namely the colocalization of the target and segmentation cue. The small effect of the control gap on task performance of both human participants and our implementation of the SS model suggests that our results cannot be fully explained by an improvement in encoding (or decoding) of target information in the SS of the stimulus facilitated by the low-level properties of the gap. Nonetheless, due to the several changes in geometry introduced in the construction of these control stimuli (see Figure 5a), it remains possible that there are some unforeseen changes in the SS of the stimuli that would allow a more elaborate version of the feedforward SS model to account for our results.

Although the effects of different kinds of target-surround similarity on contextual modulation have been widely studied for discrimination tasks using features or objects (see Whitney & Levi, 2011; Manassi & Whitney, 2018 for reviews), this has not been studied for textures (note that textures have been used to study these effects in the context of contrast perception (e.g. Wang, Heeger, & Landy, 2012; Solomon, Sperling, & Chubb, 1993). Our stimulus design allowed us to study the perceptual relevance of dissimilarity in texture properties (specifically, FAS and HOS) to target discrimination. The relation of these properties to the different stages of the SS model and of early visual processing allows us to relate or results to the model and to physiology. Previous studies using artificial textures have reported that FAS is a stronger segmentation cue than HOS, and that some HOS induce moderate and others induce weak or no segmentation (Julesz & Caelli, 1979; Victor et al., 2013; Zavitz & Baker, 2014). We found that for our naturalistic stimuli dissimilarity in FAS was a strong segmentation cue, but we did not observe clear evidence that dissimilarity in the HOS of the PS model induces segmentation in the periphery (see Figure 6). This seems also in agreement with simple inspection of our stimuli, in which the targets strongly pop out when the surround is dissimilar in FAS and HOS, but not when it is only dissimilar in HOS. The weak effect of HOS dissimilarity in peripheral vision is particularly interesting if we note that the textures with HOS dissimilarity were noticeably different under foveal inspection. It is also noteworthy that we did not observe FAS dissimilarity effects when we induced segmentation by a discontinuity between target and surround. If the surround were pooled with the target for the discontinuous condition, as the fixed pooling regions model would suggest, we would expect more similar statistics to interfere more (as was observed for our implementation of the SS model; see Figure 6), contrary to what we observed. A possible explanation for this discrepancy between our model and our data is that pooling windows are flexible, and when the surrounds are segmented from the target they are not pooled equally to when grouped together (Mareschal, Sceniak, & Shapley, 2001; Wallis et al., 2019). Finally, we showed that contextual modulation of naturalistic texture perception is strongly dependent on the naturalness of the HOS of the surround (see Figure 7), in agreement with previous perceptual (Wallis et al., 2016; Neri, 2017; Gong et al., 2018) and physiological (Guo et al., 2005; Pecka et al., 2014; Coen-Cagli et al., 2015; Ziemba et al., 2018) studies of contextual modulation.

The effects of texture structure may be informative about the mechanism of texture segmentation in the model. Human texture segmentation is a widely studied topic, and several computational models and physiological mechanisms have been proposed in the literature. Our dissimilarity results seem compatible with most of the different existing models, which is not surprising given that they can make similar predictions, and our stimuli were not designed to tell them apart. Nonetheless, our results may offer some interesting constraints on these models, and although an exhaustive analysis is out of the scope of this work, it might be useful to discuss some of the relation to three of the main biologically inspired segmentation models (Landy, 2013): the feedforward filter-rectify-filter model; the V1-based model with recurrent horizontal connections; and the multistage segmentation models with feedback.

In the classic filter-rectify-filter (FRF) kind of segmentation models, texture defined edges are detected by filtering the image with V1-like filters, rectifying the filters outputs, and then applying a second filtering stage on these outputs (Landy & Bergen, 1991; Landy, 2013; Rosenholtz, 2014). The classic version of the model uses a quadratic function for rectification, making it sensitive to local FAS for segmentation, but not to HOS in general (Landy, 2013), which seems in line with our results. Some models have been proposed to allow the FRF model to be sensitive to some HOS, such as modifying the rectifying function (Zavitz & Baker, 2013) or adding further rectification and filtering steps (Emrith, Chantler, Green, Maloney, & Clarke, 2010), but our results suggest that for naturalistic textures, these further steps may be of secondary importance.

Another class of models compatible with our dissimilarity results are the models based on recurrent contextual interactions at the level of the V1 filtering stage, that lead to differential activation at texture defined edges, allowing for segmentation and saliency (Li, 1999; Li, 2002; Robol, Grassi, & Casco, 2013; Gheorghiu, Kingdom, & Petkov, 2014), which would explain the strong segmentation effect observed for FAS dissimilarity. Interestingly, contextual interactions related to this segmentation model, such as surround suppression and surround normalization, have also been proposed to be a common computation in neural processing (Carandini & Heeger, 2012). If these contextual interactions at the level of V1 are responsible for the FAS-based segmentation, and they are also present in higher areas V2 and V4, we may expect HOS-based segmentation given the selectivity of these areas for the HOS of the PS model (Freeman et al., 2013; Okazawa et al., 2015; Ziemba et al., 2016; Okazawa et al., 2017). Nonetheless, our results showing weak segmentation for HOS dissimilarities could mean that this process of segmentation may not occur at these higher areas, or that it may be much weaker than in V1 (although see possible limitations below).

The last group of relevant models comprises the more complex and biologically inspired models involving multiple layers and recurrent feedback processing (Thielscher & Neumann, 2005; Bhatt et al., 2007; Thielscher et al., 2008; Kim, Linsley, Thakkar, & Serre, 2019). The complex nature of these models makes them difficult to analyze without actually testing them with our stimuli, although they usually use the first layer of oriented V1-like filters as the substrate of segmentation, allowing for FAS based segmentation. In addition, their feedback processing allows them to respond to more complex differences, explaining different texture segmentation results. Our results also provide an interesting experimental test to these models, namely that they should show only weak responses to the HOS explored here.

Besides the results for dissimilarity, it is less clear how our results on naturalness should be related to these models. From the discussion above, it seems that for some of these models, center and surround should not be strongly segmented if they share the same FAS. Nonetheless, it is possible that naturalness effects can emerge in some ways, particularly for the recurrent models. This could be readily tested by using implementations of these models with our stimuli as inputs. Other possible explanations for the effect of naturalness involve segmentation and contextual modulation based on probabilistic inference (Hindi Attar et al., 2007; Pecka et al., 2014; Coen-Cagli et al., 2015), although this would involve at least some extensions on the more mechanistic models described above. Finally, we note that an important limitation of our results is that although the selectivity of areas V2 and V4 to naturalistic HOS is well established (Freeman et al., 2013; Okazawa et al., 2015; Ziemba et al., 2016; Okazawa et al., 2017), this has not been tested for stimuli with different HOS but matched FAS as those used in this work. Furthermore, the space of PS statistics is high dimensional, and it is possible that other dissimilarities in HOS produce strong segmentation (although note that the textures with dissimilar HOS look considerably different under foveal inspection). Indeed, previous work with artificial textures shows that selectivity for other simpler HOS that can support texture segmentation (Victor et al., 2013) emerges primarily in V2 (Yu, Schmid, & Victor, 2015). Therefore, a more exhaustive exploration of the capacity of naturalistic HOS to induce segmentation would be needed to better understand their role in segmentation, as well as possible contributions from higher visual areas.

What neural mechanisms might underlie the contextual modulation we observe? One candidate is V1 surround suppression, which appears linked to our experimental results in several ways: both strongly depend on FAS similarity (Cavanaugh, Bair, & Movshon, 2002) and on segmentation cues (Coen-Cagli et al., 2015), and it has been proposed that V1 surround suppression underlies perceptual surround suppression (Zenger-Landolt & Heeger, 2003; Carandini & Heeger, 2012), which affects texture perception (Chubb, Sperling, & Solomon, 1989; McDonald & Tadmor, 2006; Wang et al., 2012) and is relatively strong in peripheral vision (Xing & Heeger, 2000; Petrov, Carandini, & McKee, 2005). In addition, we showed that contextual modulation of naturalistic texture perception is tuned to the naturalness of the HOS, in agreement with previous perceptual (Wallis et al., 2016; Neri, 2017; Gong et al., 2018) and physiological (Guo et al., 2005; Pecka et al., 2014; Coen-Cagli et al., 2015; Ziemba et al., 2018) studies in contextual modulation. This too could reflect V1 surround suppression, which has been shown to be reduced for scrambled surrounds (i.e. lacking natural HOS) compared to natural images in V1 (Guo et al., 2005; Pecka et al., 2014; Coen-Cagli et al., 2015; although unpublished recordings indicate this might not be the case for naturalistic textures Ziemba, Tim Oleskiw, Perez, Simoncelli, & Movshon, 2017). Overall, our experimental results on contextual modulation and segmentation appear consistent with flexible V1 surround suppression (Coen-Cagli et al., 2015), in which suppression strength is reduced when center and surround are inferred to be segmented on the basis of image statistics. Furthermore, as discussed above, this recurrent process of contextual modulation in V1 is also related to some segmentation models (Li, 1999; Li, 2002; Schmid, 2008; Robol et al., 2013; Gheorghiu et al., 2014), and it could also partly explain the segmentation effects we observed. Following the proposed matching of physiology and the SS model (see Figure 1), this process would act after the filtering stage of the model, prior to computing the SS of the texture features.

Another possible mechanism relevant to our results is facilitation. For example, one possible contributor is surround facilitation at the level of V2, observed in texture stimuli similar to ours, in which the response of V2 neurons to a texture patch can be enhanced by naturalistic texture surrounds outside their receptive fields (Ziemba et al., 2018). Following the parallel between physiology and the SS model (see Figure 1), this mechanism would act over the output of the SS computation. After the SS of the different image regions are computed, naturalistic surrounds would facilitate the output of the SS computing units corresponding to the target. Although not directly tested in this previous study, this facilitation mechanism could be stronger for scrambled targets than for naturalistic targets, reducing the difference in responses between the two kinds of targets when naturalistic surrounds are included. Thus, this reduced difference between the SS of the two kinds of targets would result in reduced target discriminability. If this is a relevant mechanism in our task, then our results would suggest that V2 surround facilitation is reduced by target surround segmentation, and that it is relatively stronger for phase scrambled targets than for naturalistic targets, which could be readily tested experimentally.

Another mechanism that may contribute to our results is pooling over flexible windows shaped by segmentation, as proposed in studies of natural scene perception in peripheral vision (Wallis et al., 2019) and orientation discrimination in central vision (Mareschal et al., 2001). Flexible surround suppression, facilitation, and flexible pooling windows could therefore be integrated at the corresponding stages of the SS model, leading to a broader framework within which to interpret our results and guide further studies of peripheral vision.

Although the results discussed so far appeared consistent with perceptual crowding, we did not observe a clear inwards-outwards asymmetry as is often reported for crowding (Petrov et al., 2007; see Figures 8, 9). One possible interpretation for this result is that the processes dominating our contextual modulation effect may not be the same as those in letter crowding. Nonetheless, another possible interpretation is that our contextual modulation phenomenon is produced by the same mechanisms as letter crowding, but that these mechanisms affect texture perception in our task differently from commonly used stimuli such as letters and vernier. For example, it has been shown that inward and outward flankers have different relative weight for different crowding processes and different crowding tasks (Chastain, 1982; Strasburger & Malania, 2013; Strasburger, 2019), and also that classical inward-outward asymmetry can be reversed by biasing attention towards the fovea (Petrov & Meleshkevich, 2011b). Therefore, it is possible that if the different subprocesses of crowding have different relative effects on textures than on letters, this may lead to a different overall inwards-outwards asymmetry. In line with this, we speculate that the large variability we observed across participants in the sensitivity to the different experimental conditions, such as surround position, reflects a variability in the relevance of the different underlying processes (whether the same as in classical crowding or not), which is also consistent with other crowding studies (Kooi et al., 1994; Petrov & Meleshkevich, 2011a; Wallace et al., 2013; Lev & Polat, 2015). Although our results do not allow to tell whether our contextual modulation phenomenon is different from letter crowding, or if it involves the same mechanisms as crowding but affecting textures differently, they point to the need of further studies on the relation between contextual modulation of textures and the phenomenon of crowding, which is frequently described as objects undergoing “forced texture processing” (Rosenholtz, 2016). This would also be relevant, for example, to previous work studying crowding in natural scenes (Wallis & Bex, 2012; Gong et al., 2018), which according to this line of reasoning might also have measured, to an unknown degree, other contextual modulation processes affecting texture perception. As explained above, our work points to additional processes such as flexible surround suppression and facilitation whose relation to crowding is uncertain and which may be of particularly high relevance to texture contextual modulation. Finally, it is worth noting that some of the tasks most associated with peripheral vision such as scene perception (Ehinger & Rosenholtz, 2016; Brady, Shafer-Skelton, & Alvarez, 2017; Groen, Silson, & Baker, 2017), guidance of eye movements (Parkhurst & Niebur, 2004; Frey, König, & Einhäuser, 2007; Schmid & Victor, 2014) and the control of body movement (Brandt, Dichgans, & Koenig, 1973; Bardy, Warren, & Kay, 1999; Berencsi, Ishihara, & Imanaka, 2005) have been proposed to use texture as a major source of information (Harrington et al., 1985; Sinai, Krebs, Darken, Rowland, & McCarley, 1999; Parkhurst & Niebur, 2004; Frey et al., 2007; Schmid & Victor, 2014; Brady et al., 2017; Groen et al., 2017; Ehinger & Rosenholtz, 2016). Therefore, understanding the role of contextual modulation on texture perception in the periphery may be an important step for understanding of the limitations of peripheral vision in natural behavior.

Keywords: texture, naturalistic, contextual modulation, peripheral vision

Acknowledgments

The authors thank Corey Ziemba and Adrien Doerig for useful discussions on an earlier version of this manuscript. We also thank Adam Kohn for comments on an earlier version of this manuscript. This work was funded by the studentships awarded to D. Herrera and L. Gómez-Sena by the Comisión Académica de Posgrados, UdelaR, Uruguay, and by travelships awarded to D. Herrera by PEDECIBA, UdelaR, Uruguay, and CSIC, Uruguay. Ruben Coen-Cagli was supported in part by National Institutes of Health grant EY031166.

Commercial relationships: none.

Corresponding author: Daniel Herrera-Esposito.
Email: dherrera@fcien.edu.uy.
Address: Laboratorio de Neurociencias, Facultad de Ciencias, Universidad de la República, Avenida Gral. Flores 2125, Montevideo, Uruguay.

*RCC and LGS contributed equally to this work.

References

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

Banks, W. P., Larson, D. W., & Prinzmetal, W. (1979). Asymmetry of visual interference. *Perception & Psychophysics, 25*(6), 447–456.

Bardy, B. G., Warren, W. H., & Kay, B. A. (1999). The role of central and peripheral vision in postural control during walking. *Perception & Psychophysics, 61*(7), 1356–1368.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. arXiv:1506.04967 [stat]. arXiv: 1506.04967. Retrieved April 21, 2020. Available at: http://arxiv.org/abs/1506.04967.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., . . . Dai, B. et al. (2019). Lme4: Linear Mixed-Effects Models using 'Eigen' and S4. Retrieved April 21, 2020. Available at: https://rdrr.io/cran/lme4/.

Berencsi, A., Ishihara, M., & Imanaka, K. (2005). The functional role of central and peripheral vision in the control of posture. *Human Movement Science. Neural, Cognitive and Dynamic Perspectives of Motor Control, 24*(5), 689–709.

Bergen, J. R., & Landy, M. S. (1991). Computational Modeling of Visual Texture Segregation. In M. Landy, & J. A. Movshon (Eds.), *Computational Models of Visual Processing* (pp. 253–271). Cambridge, MA: MIT Press.

Bhatt, R., Carpenter, G. A., & Grossberg, S. (2007). Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research, 47*(25), 3173–3211.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1160–1176.

Brandt, T., Dichgans, J., & Koenig, E. (1973). Differential effects of central versus peripheral vision on egocentric and exocentric motion perception. *Experimental Brain Research, 16*(5), 476–491.

Burghouts, G. J., & Geusebroek, J.-M. (2009). Material-specific adaptation of color invariant features. *Pattern Recognition Letters, 30*(3), 306–313.

Cant, J. S., & Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *Journal of Neuroscience, 32*(22), 7685–7700.

Cant, J. S., Large, M.-E., McCall, L., & Goodale, M. A. (2008). *Independent Processing of Form, Colour, and Texture in Object Perception: Perception*. London, England: SAGE Publication Sage UK.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*(1), 51–62.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology, 88*(5), 2547–2556.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010). Separate channels for processing form, texture, and color: Evidence from fMRI adaptation and visual object agnosia. *Cerebral Cortex, 20*(10), 2319–2332.

Chastain, G. (1982). Confusability and interference between members of parafoveal letter pairs. *Perception & Psychophysics, 32*(6), 576–580.

Chubb, C., Sperling, G., & Solomon, J. A. (1989). Texture interactions determine perceived contrast. *Proceedings of the National Academy of Sciences, 86*(23), 9631–9635.

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience, 18*(11), 1648–1655.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences, 20*(5), 324–335.

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.

Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision, 16*(2), 13–13.

Emrith, K., Chantler, M. J., Green, P. R., Maloney, L. T., & Clarke, A. D. F. (2010). Measuring perceived differences in surface texture due to changes in higher order statistics. *JOSA A, 27*(5), 1232–1244.

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision, 9*(6), 18.

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation ex plain properties of visual crowding. *Psychological Review, 124*(4), 483–504.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974–981.

Frey, H.-P., König, P., & Einhäuser, W. (2007). The role of first- and second-order stimulus features for human overt attention. *Perception & Psychophysics, 69*(2), 153–161.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Simon, N., & Qian, J. (2019). Glmnet: Lasso and elastic-net regularized generalized linear models. Retrieved April 21, 2020. Available at: https://rdrr.io/cran/glmnet/.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models. Google-Books-ID: c9xLKzZWoZ4C*. Cambridge, MA: Cambridge University Press.

Gheorghiu, E., Kingdom, F. A. A., & Petkov, N. (2014). Contextual modulation as de-texturizer. Vision Research. *The Function of Contextual Modulation, 104*, 12–23.

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

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.

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1714), 20160102.

Guo, K., Robertson, R. G., Mahmoodi, S., & Young, M. P. (2005). Centre-surround interactions in response to natural scene stimulation in the primary visual cortex. *European Journal of Neuroscience, 21*(2), 536–548.

Harrington, T. L., Harrington, M. K., Quon, D., Atkinson, R., Cairns, R., & Kline, K. (1985). Perception of orientation of motion as affected by change in divergence of texture, change in size, and in velocity. *Perceptual and Motor Skills, 61*(3), 875–886.

Hermundstad, A. M., Briguglio, J. J., Conte, M. M., Victor, J. D., Balasubramanian, V., & Tkačik, G. (2014). Variance predicts salience in central sensory processing. *eLife, 3*, e03722.

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.

Hindi Attar, C., Hamburger, K., Rosenholtz, R., Götzl, H., & Spillmann, L. (2007). Uniform versus random orientation in fading and filling-in. *Vision Research, 47*(24), 3041–3051.

Eaton, J. W., Bateman, D., Hauberg, S., & Wehbring, R. (2015). GNU Octave version 4.0.0 manual: A high-level interactive language for numerical computations.

Julesz, B. (1962). Visual pattern discrimination. *IRE Transactions on Information Theory, 8*(2), 84–92.

Julesz, B., Gilbert, E. N., & Victor, J. D. (1978). Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics, 31*(3), 137–140.

Julesz, B., & Caelli, T. (1979). On the Limits of Fourier decompositions in visual texture perception. *Perception, 8*, 69–73.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. In *International Conference on Learning Representations*, Retrieved June 26, 2020. Available at: https://openreview.net/forum?id=HJxrVA4FDS.

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? *Perception, 36*, S14.

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.

Landy, M. S. (2013). Texture analysis and perception. In J.S. Werner, & L.M. Chalupa (Eds.), *The new visual neurosciences* (pp. 639–652). Cambridge, MA: MIT Press.

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research, 31*(4), 679–691.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1265–1278.

Lev, M., & Polat, U. (2015). Space and time in masking and crowding. *Journal of Vision, 15*(13), 10.

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

Li, Z. (1999). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences, 96*(18), 10530–10535.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16.

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.

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

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 uncrowd ing. *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.

Mareschal, I., Sceniak, M. P., & Shapley, R. M. (2001). Contextual influences on orientation discrimination: Binding local and global cues. *Vision Research, 41*(15), 1915–1930.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron, 71*(5), 926–940.

McDonald, J. S., & Tadmor, Y. (2006). The perceived contrast of texture patches embedded in natural images. *Vision Research, 46*(19), 3098–3104.

McWalter, R., & McDermott, J. H. (2018). Adaptive and selective time averaging of auditory scenes. *Current Biology, 28*(9), 1405–1418.e10.

Meinecke, C., & Kehrer, L. (1994). Peripheral and foveal segmentation of angle textures. *Perception & Psychophysics, 56*(3), 326–334.

Morikawa, K. (2000). Central performance drop in texture segmentation: The role of spatial and temporal factors. *Vision Research, 40*(25), 3517–3526.

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.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences, 112*(4), E351–E360.

Okazawa, G., Tajima, S., & Komatsu, H. (2017). Gradual development of visual texture-selective properties between macaque areas V2 and V4. *Cerebral Cortex, 27*(10), 4867–4880.

Overvliet, K. E., & Sayim, B. (2016). Perceptual grouping determines haptic contextual modulation.

Vision Research. *Quantitative Approaches in Gestalt Perception, 126*, 52–58.

Paradiso, M. A., & Nakayama, K. (1991). Brightness perception and filling-in. *Vision Research, 31*(7), 1221–1236.

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.

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19*(3), 783–789.

Pecka, M., Han, Y., Sader, E., & Mrsic-Flogel, T. D. (2014). Experience-dependent specialization of receptive field surround for selective coding of natural scenes. *Neuron, 84*(2), 457–469.

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

Petrov, Y., Carandini, M., & McKee, S. (2005). Two distinct mechanisms of suppression in human vision. *Journal of Neuroscience, 25*(38), 8704–8707.

Petrov, Y., & Meleshkevich, O. (2011a). Asymmetries and idiosyncratic hot spots in crowding. *Vision Research, 51*(10), 1117–1123.

Petrov, Y., & Meleshkevich, O. (2011b). Locus of spatial attention determines inward–outward anisotropy in crowding. *Journal of Vision, 11*(4), 1.

Petrov, Y., Popple, A. V., & McKee, S. P. (2007). Crowding and surround suppression: Not to be confused. *Journal of Vision, 7*(2), 12.

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

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.

Qiu, C., Kersten, D., & Olman, C. A. (2013). Segmentation decreases the magnitude of the tilt illusion. *Journal of Vision, 13*(13), 19.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria, https://elifesciences.org/articles/42512, https://www.r-bloggers.com/2018/06/its-easy-to-cite-and-reference-r/.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Robinson, D., & Hayes, Alex. (2018). *Broom: Convert Statistical Objects into Tidy Tibbles in broom: Convert Statistical Analysis Objects into Tidy Tibbles*, Retrieved April 21, 2020. Available at: https://rdrr.io/cran/broom/man/broom.html.

Robol, V., Grassi, M., & Casco, C. (2013). Contextual influences in texture-segmentation: Distinct effects from elements along the edge and in the texture-region. *Vision Research, 88*, 1–8.

Rosen, S., Chakravarthi, R., & Pelli, D. G. (2014). The Bouma law of crowding, revised: Critical spacing is equal across parts, not objects. *Journal of Vision, 14*(6), 10.

Rosenholtz, R. (2014). Texture perception. In *The Oxford Handbook of Perceptual Organization*, https://doi.org/10.1093/oxfordhb/9780199686858.013.058.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*(1), 437–457.

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.

Saarela, T. P., & Herzog, M. H. (2009). Size tuning and contextual modulation of backward contrast masking. *Journal of Vision, 9*(11), 21.

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., Westheimer, G., & Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in vernier acuity. *Journal of Vision, 8*(8), 12.

Schade, U., & Meinecke, C. (2009). Spatial distance between target and irrelevant patch modulates detection in a texture segmentation task. *Spatial Vision, 22*(6), 511–527.

Schade, U., & Meinecke, C. (2011). Texture segmentation: Do the processing units on the saliency map increase with eccentricity? *Vision Research, 51*(1), 1–12.

Schmid, A. M. (2008). The processing of feature discontinuities for different cue types in primary visual cortex. *Brain Research, 1238*, 59–74.

Schmid, A. M., & Victor, J. D. (2014). Possible functions of contextual modulations and receptive field nonlinearities: Pop-out and texture segmentation. *Vision Research. The Function of Contextual Modulation, 104*, 57–67.

Simoncelli, E., Freeman, W., Adelson, E., & Heeger, D. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory, 38*(2), 587–607.

Sinai, M., Krebs, W., Darken, R., Rowland, J., & McCarley, J. (1999). Egocentric distance perception in a virtual environment using a perceptual matching task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43*(22), 1256–1260.

Solomon, J. A., Sperling, G., & Chubb, C. (1993). The lateral inhibition of perceived contrast is indif ferent to on-center/off-center segregation, but specific to orientation. *Vision Research, 33*(18), 2671–2683.

Strasburger, H. (2019). Seven myths on crowding and peripheral vision. *PeerJ Preprints*, https://doi.org/10.7287/peerj.preprints.27353v4.

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

Stürzel, F., & Spillmann, L. (2001). Texture fading correlates with stimulus salience. *Vision Research, 41*(23), 2969–2977.

Thielscher, A., Kölle, M., Neumann, H., Spitzer, M., & Grön, G. (2008). Texture segmentation in human perception: A combined modeling and fMRI study. *Neuroscience, 151*(3), 730–736.

Thielscher, A., & Neumann, H. (2005). Neural mechanisms of human texture processing: Texture boundary detection and visual search. *Spatial Vision, 18*(2), 227–257.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*(17), 2503–2522.

Vancleef, K., Putzeys, T., Gheorghiu, E., Sassi, M., Machilsen, B., & Wagemans, J. (2013). Spatial arrangement in texture discrimination and texture segregation. *i-Perception, 4*(1), 36–52.

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with s (4th ed.). *Statistics and Computing*. New York: Springer-Verlag.

Vergeer, M. L. T., & van Lier, R. (2007). Grouping effects in flash-induced perceptual fading. *Perception, 36*(7), 1036–1042.

Victor, J. D. (1994). Images, statistics, and textures: Implications of triple correlation uniqueness for texture statistics and the Julesz conjecture: *Comment. JOSA A, 11*(5), 1680–1684.

Victor, J. D., Conte, M. M., & Chubb, C. F. (2017). Textures as probes of visual processing. *Annual Review of Vision Science, 3*(1), 275–296.

Victor, J. D., Thengone, D. J., & Conte, M. M. (2013). Perception of second- and third-order orientation signals and their interactions. *Journal of Vision, 13*(4), 21–21.

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. S. A., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision, 16*(2), 4.

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.

Wallis, T. S. A., & Bex, Peter J.. (2012). Image correlates of crowding in natural scenes. *Journal of Vision, 12*(6), 1–19.

Wang, H. X., Heeger, D. J., & Landy, M. S. (2012). Responses to second-order texture modulations undergo surround suppression. *Vision Research, 62*, 192–200.

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.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis. Google-Books-ID: XgFkDAAAQBAJ*. New York, NY: Springer.

Wickham, H., Henry, L., & RStudio. (2018). Tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions.

Wickham, H., François, R., Henry, L., & Müller, K. (2018). Dplyr: A grammar of data manipulation.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*(3), 671–684.

Xie, Y. (2015). *Dynamic Documents with r and knitr. Google-Books-ID: lpTYCQAAQBAJ*. Boca Raton, FL: CRC Press.

Xing, J., & Heeger, D. J. (2000). Center-surround interactions in foveal and peripheral vision. *Vision Research, 40*(22), 3065–3072.

Yu, Y., Schmid, A. M., & Victor, J. D. (2015). Visual processing of informative multipoint correlations arises primarily in V2. *eLife, 4*, e06604.

Zavitz, E., & Baker, C. L. (2013). Texture sparseness, but not local phase structure, impairs second order segmentation. *Vision Research, 91*, 45–55.

Zavitz, E., & Baker, C. L. (2014). Higher order image structure enables boundary segmentation in the absence of luminance or contrast cues. *Journal of Vision, 14*(4), 14.

Zenger-Landolt, B., & Heeger, D. J. (2003). Response suppression in V1 agrees with psychophysics of surround masking. *Journal of Neuroscience, 23*(17), 6884–6893.

Zenger-Landolt, B., & Koch, C. (2001). Flanker effects in peripheral contrast discrimination—psychophysics and modeling. *Vision Research, 41*(27), 3663–3675.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences, 113*(22), E3140–E3149.

Ziemba, C. M., Freeman, J., Simoncelli, E. P., & Movshon, J. A. (2018). Contextual modulation of sensitivity to naturalistic image structure in macaque V2. *Journal of Neurophysiology, 120*(2), 409–420.

Ziemba, C. M., Tim, Oleskiw, Perez, R. K., Simoncelli, E. P., & Movshon, J. A. (2017). Selectivity of contextual modulation in macaque V1 and V2. *Annual Meeting, Neuroscience*. Retrieved April 21, 2020. Available at: https://www.cns.nyu.edu/∼lcv/pubs/makeAbs.php?loc=Ziemba17b.

Supplementary material

S1 Converting log-odds ratios to proportion differences

The goal of our statistical analysis is to estimate the effect on performance of changing a given condition, and testing whether the effect is significantly different from 0. All the conditions we analyze can be taken to vary in a binary way, being present or absent (e.g. discontinuity, surround presence, HOS dissimilarity, etc), and thus we frame the analysis as estimating the effect of having that condition present. Also, it is expected that the effect of a given condition can vary across people and across textures. We are therefore interested in estimating the mean effect of that condition across the population of textures and participants, and taking the variability into account when estimating whether it is significantly different from 0.

For this goal we fit a generalized linear mixed model (GLMM) of the binomial family (equivalent to logistic regression) [27]. This procedure estimates the mean effect of the presence of a given condition A on the probability of success, and also estimates the variability of this effect across the population from which textures and participants were sampled. Taking condition A (e.g. discontinuity) to be a variable with two possible values, A = 1 (A present) and A = 0 (A absent), the model relating A to task performance (probability of success, P) for a given participant s with a given texture t is the following:

$$P(Success|A = a, Texture = t, Participant = s) = \frac{1}{1 + e^{-l(a,t,s)}} \tag{S1}$$

$$l(A = a, Texture = t, Participant = s) = a\beta_A^{t,s} + \beta_0^{t,s} \tag{S2}$$

where $\beta_A^{t,s}$ is the effect of A and $\beta_0^{t,s}$ is the offset for participant s and texture t. Here $\beta_A^{t,s}$ is a sample from a stochastic variable given by $\beta_A^{t,s} = \beta_A + Z_A^t + Z_A^s$. Here β_A is the mean effect of condition A (or fixed effect), and $Z_A^t \sim \mathcal{N}(0, \sigma_{A:t}^2)$ and $Z_A^s \sim \mathcal{N}(0, \sigma_{A:s}^2)$ are samples from random variables corresponding to the variability of the effect across textures and participants respectively (the random effects). The random effects are characterized by their standard deviations $\sigma_{A:t}$ and $\sigma_{A:s}$. Similarly, $\beta_0^{t,s} = \beta_0 + Z_0^t + Z_0^s$, with $Z_0^t \sim \mathcal{N}(0, \sigma_{0:t}^2)$ and $Z_0^s \sim \mathcal{N}(0, \sigma_{0:s}^2)$. We note that all participants under a given texture share the same value of Z_A^t and Z_0^t, meaning that each texture has its own characteristic effect of A and offset. Fitting a GLMM model to an experiment in which condition A is varied estimates all of the above parameters. The discussions in the text are based on the estimates of the fixed effects for the experimental manipulations, which is the estimated mean effect across textures and subjects.

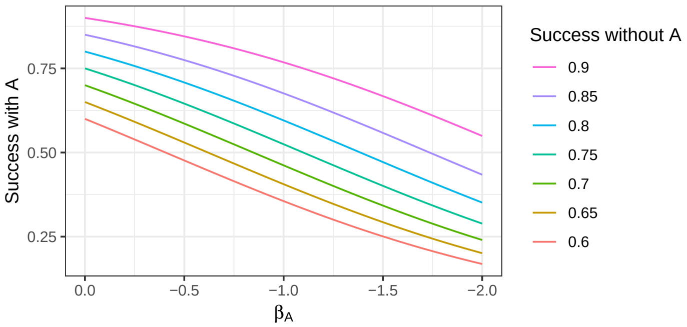


Figure S1: Reference plot for converting LOR (β_A) into proportions of correct answers. Each line indicates a different initial probability of success without condition A (e.g. discontinuity, naturalness, etc). The lines then show the probability of success with condition A for the different values of β_A.

As mentioned, the effects estimated by the model are the β_A coefficients in the linear equation l inside the logistic function. These effects do not directly express the difference in success probabilities between the conditions, they express the log-odds-ratio (LOR) between the conditions, which is a measure related to the difference in probability of success. The odds of success for a given condition can be defined as $odds = \frac{p_{success}}{p_{fail}}$, and it can be converted to probability of success. The odds-ratio is, as the name suggests, the ratio of the odds of two conditions, and it is a measure of the difference in their success probabilities. For example, a way of expressing the effect of condition A on task performance is as the odds-ratio between the condition with A present (A=1) and with A absent (A=0), or $OR_A = \frac{odds_{A=1}}{odds_{A=0}}$. If the presence of A improves performance, OR_A will be larger than 1, and if it hinders performance, OR_A will be between 1 and 0. The log-odds-ratio (LOR) is simply the logarithm of the odds-ratio, and so the parameters estimated by the model can also be understood as $\beta_A = \log OR_A$. The LOR has different advantages as a measure of changes in probability of succes. For one, it is unbounded (can go from $-\infty$ to $+\infty$) and symmetric around 0 (the LOR of a change in probability from p_1 to p_2 is the opposite of the LOR of a change from p_2 to p_1). But also, the LOR has the advantage that it reflects an intuitive aspect of the relevance in a change in probability. Intuitively, the relevance of a change in probability depends on the specific probability values, and this is readily reflected in the LOR, but not in the raw probability changes. For example, a change in p from 0.9 to 0.99 is intuitively more important than a change from 0.5 to 0.59. While the change in probability is the same in both cases, the LOR between the conditions are around 2.40 and 0.36 respectively.

But the above mentioned advantage of the LOR means that a given LOR does not uniquely determine a change in probability, since the change in probability will depend on the specific probability values. Therefore we express in the article the estimated effects as the LOR (the βs). To help get intuition of the magnitude of those effects in terms of probability, Figure S1 shows how to convert LOR to changes in probability. To see how a given LOR translates to a difference in success probability for an experimental manipulation, select an initial probability (a given line) and see what the probability is at the selected LOR. Those would be the probabilities of success without and with condition A present, respectively. Note that a given LOR gives different changes in probability for different initial probabilities.

S2 Experiment 1

In experiment 1 we expected the width of the gap to be an important factor in determining the effect of discontinuity because of two opposing factors. For one, since the gap is produced by shrinking the target, a larger gap implies a smaller target for the discontinuous condition. This could lead to a reduction in performance for the discontinuous condition, and potentially mask the effect of segmentation. On the other hand, a smaller gap can be less visible (particularly with the transparency gradients at the borders of the textures), leading to weaker segmentation and thus also leading to a reduced performance for the discontinuous condition. Therefore, we expected the effect of discontinuity to be maximal at some intermediate gap width. We chose to run the experiment using two different gap widths, 0.3° and 0.6°, which seemed to be sufficiently visible but not too large. In experiment 1 we report the result of discontinuity for the gap with 0.3° width. The gap with 0.6° width that showed an effect in the same direction, albeit smaller and non significant ($\beta_{Discont} = 0.27$, ci = $\left[-8 \times 10^{-3}, 0.54\right]$, p = 0.06, Figure S2c). Thus, although the points in the discussion remain unchanged, we note that the experimental effect depends on this relevant parameter.

Also, in experiment 1 we found that texture surrounds impair task performance and that target-surround segmentation can reduce this impairment. Thus, we tested whether segmentation completely recovered task performance. For this we compared task performance for the unsurrounded target, and

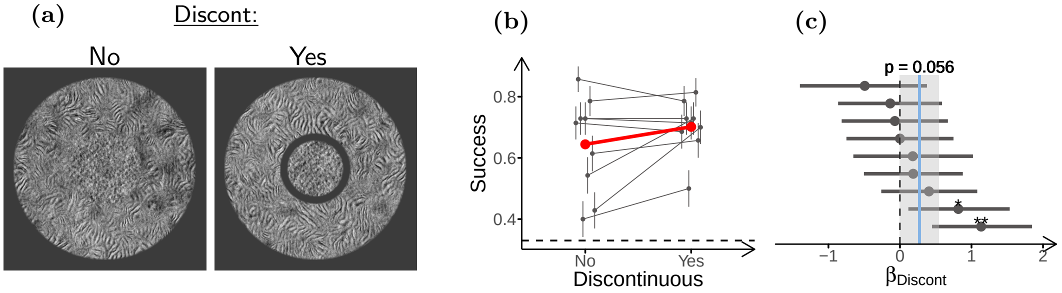


Figure S2: The effect of discontinuity depends on the size of the gap. **(a)** Stimulus configurations used in the experiment (only scrambled targets shown). Left: Continuous stimulus (larger target size), Right: Discontinuous stimulus (smaller target size) with a gap of 0.6°. **(b)** Task performance for the two conditions. **(c)** LOR for discontinuity ($\beta_{Discont}$), estimated from the performance data in **(b)**. Participants (n=8) performed 70 trials in each condition. The plots in the figure use the same conventions as Figure 3.

for stimuli with surround texture but separated from the targets by a gap (Figure S3a). For the discontinuous condition we pooled the data from the two gap sizes (0.3° and 0.6°). We note that the discontinuous and unsurrounded stimuli had the same target size, only differing in the presence of the discontinuous surround. We found that the discontinuous surrounds still impaired performance ($\beta_{Surr} = -0.40$, ci $= [-0.68, -0.15]$, p $= 5 \times 10^{-3}$), showing that segmentation did not completely remove contextual modulation.

We note that all the conditions mentioned in the section corresponding to experiment 1 (the ones for testing the effect of flanker and the effect of discontinuity), as well as those presented here, were presented in the same experimental session. That is, participants shown in Figures 3 and 4 and in this section are the same individuals (except for a missing participant in Figure 3, which was not presented the unsurrounded due to an error).

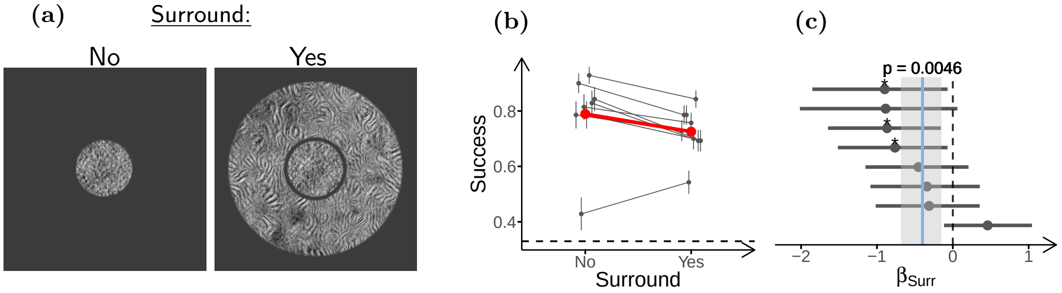


Figure S3: Discontinuous surrounds generate contextual modulation. **(a)** Stimulus configurations used in the experiment (only scrambled targets shown). Left: Target without surround, Right: Target with discontinuous surround. **(b)** Task performance for the two conditions. **(c)** LOR for the presence of the discontinuous surround (β_{Surr}), estimated from the performance data in **(b)**. 8 participants performed 70 trials in the unsurrounded condition, and 140 in the surrounded condition. The plots in the figure use the same conventions as Figure 3.

S3 Experiment 2

In experiment 2 we induced continuity by changing target shape from disk-target to split-target (as described in the main text). Although the total area of target texture is roughly the same for the two shapes, it is possible that part of the observed effect could be due to target shape rather than discontinuity. To test for this we included control stimuli for texture T1, consisting of the targets without surrounds, and the targets with surrounds but without a gap, for both target shapes.

First we estimated the effect of target shape by comparing performances for the two targets without surrounds (Figure S4a). The performance for the split-target was slighly worse than for the disk target (Figure S4b), although the effect was small and non-significant ($\beta_{Split} = -0.16$, ci $= [-0.42, 0.11]$, p $= 0.22$, Figure S4c).

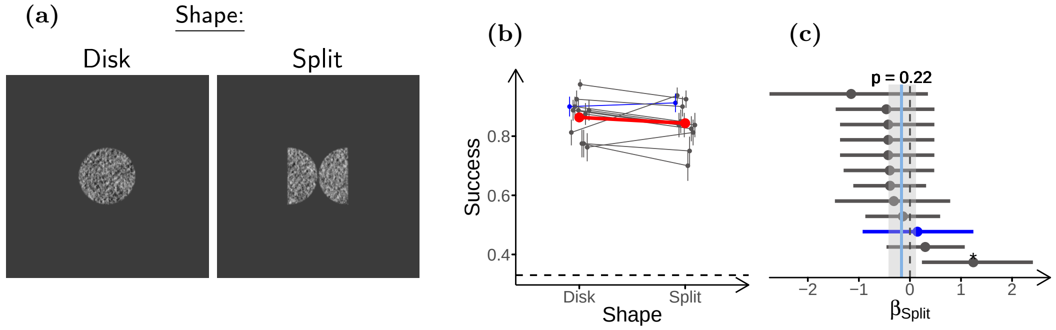


Figure S4: Target shape does not affect performance for targets without surrounds. **(a)** Stimulus configurations used in the experiment. Left: Non-split target (Disk-target) without surround, Right: Split-target without surround. **(b)** Task performance for the two conditions. **(c)** LOR for the splitting of the target (β_{Split}), estimated from the data in **(b)**. Participants (n=12) performed 80 trials per condition. The plots in the figure use the same conventions as Figure 3.

Then we estimated the effect of target shape and whether it explains the effect of discontinuity. For this we fitted a model with terms for target shape (β_{Split}), the presence of the gap (β_{Gap}), and discontinuity ($\beta_{Discont}$) to the stimuli shown in S5a, which include targets of both shapes with surrounds, and in presence and absence of a gap around the target. This analysis also allows to estimate the effect of having a gap around the target after accounting for the discontinuity effect (e.g. effects arising from the low level properties of the gap, or for spatial cueing to the target location within the stimulus). Interestingly, both the effect of target shape ($\beta_{Split} = -0.09$, ci $= [-0.31, 0.14]$, p $= 0.43$, Figure S5c) and of the presence of the gap ($\beta_{Gap} = 0.04$, ci $= [-0.15, 0.22]$, p $= 0.71$, Figure S5d) were close to 0 and non significant, while the effect of discontinuity had a similar magnitude as to that estimated in the main text ($\beta_{Discont} = 0.43$, ci $= [0.14, 0.73]$, p $= 5 \times 10^{-3}$, Figure S5e). This can be seen in the raw performances (Figure S5b), where in the absence of a gap both target shapes have similar performances (thus showing a small effect of target-shape), but in the presence of a gap there is a difference in performance (since one shape is discontinuous with the surround while the other is not). The small effect of the gap can be seen in the little difference between the conditions with and without a gap for the split target, suggesting that the gap does not have a strong effect if it does not induce segmentation. This shows that the improvement in performance described for experiment 2 in the main text comes mainly from discontinuity, rather than target shape or other low level factors that accompany the presence of the gap.

S4 Experiment 3

In experiment 3, when estimating the effect of target-surround dissimilarity for the discontinuous and continuous conditions separately, there appears to be a considerable change in the effect of FAS dissimilarity. To verify that this change is significant we fitted a GLMM to all the data shown in Figure 6, with parameters for FAS dissimilarity (β_{FAS}), HOS dissimilarity (β_{HOS}), discontinuity ($\beta_{Discont}$), and the interactions between discontinuity and dissimilarity ($\beta_{FAS:Discont}$ and $\beta_{HOS:Discont}$). In

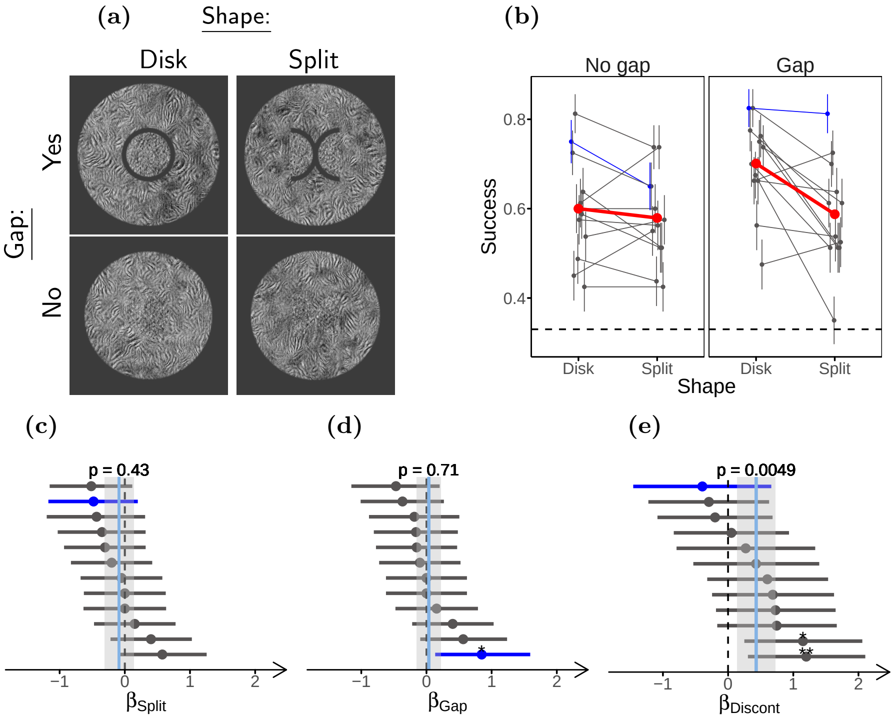


Figure S5: Target shape does not affect performance for targets with surrounds. **(a)** Stimulus configurations used in the experiment. Left: Disk-target shape, Right: Split-target shape. Top: Stimuli with a gap. Bottom: Stimuli without a gap. **(b)** Task performance for the four conditions, with the conditions without a gap in the left panel and the conditions with a gap around the target in the right panel. **(c)-(e)** LOR for, respectively, the splitting of the target (β_{Split}), the presence of the gap (β_{Gap}), and discontinuity ($\beta_{Discont}$), estimated from the data in **(b)**. Participants (n=12) participants performed 80 trials per condition. The plots in the figure use the same conventions as Figure 3.

this model, β_{FAS} and β_{HOS} are the effect of dissimilarity for the continuous condition, $\beta_{Discont}$ is the effect of discontinuity for the condition with no target-surround dissimilarity, and the interaction terms represent how the effect of dissimilarity changes for the discontinuous condition (or equivalently, how the effect of discontinuity changes for the dissimilar surrounds). We note that due to convergence issues in computing the confidence intervals of the parameters, we excluded the random effects for $\beta_{FAS:Discont}$ from the model, although these were estimated by the model to be close to 0 when fitting the full model, and both the point estimates of the parameters and their corresponding p-values were practically unchanged when including these random effects.

As expected, the full model fit shows a strong and negative effect for the FAS-discontinuity interaction ($\beta_{FAS:Discont} = -0.78$, ci $= [-0.95, -0.61]$, p $= 0.$, Figure S6c) indicating that the effect of FAS dissimilarity is reduced (and practically canceled) when the stimulus is discontinuous. The interaction term between the HOS and discontinuity on the other hand was close to 0 and non significant ($\beta_{HOS:Discont} = 0.07$, ci $= [-0.35, 0.49]$, p $= 0.66$, Figure S6e) and showed considerable variability across textures.

We also estimated the interaction effects for the simulated observers. In line with the similarity between the effects for continuous and discontinuous conditions in Figure 6, the interaction between FAS dissimilarity and discontinuity was close to 0 and non significant ($\beta_{FAS:Discont} = 0.02$, ci $=$ $[-0.02, 0.06]$, p $= 0.38$, Figure S6h). Thus, this confirms the failure of the model to capture the experimental results. Interestingly, the model showed a small but significant interaction between HOS and discontinuity ($\beta_{HOS:Discont} = 0.11$, ci $= [0.07, 0.15]$, p $= 8 \times 10^{-4}$, Figure S6j). This interaction may result from the change in the amount of texture around the target for the discontinuous condition, since the surround texture in the gap is removed, and the magnitude effect will depend on whether HOS are matched or not.

Also, we note that for texture T1 experiment 3 was done using two different widths which, were 0.5° and 1°. Due to the small difference observed between gap sizes in this experiment and preliminary data, we resolved to continue with only the smaller gap for the rest of the experiments. To verify whether the effect of gap width was negligible, we fitted a GLMM to the data with terms for gap width and found them all to be small and non-significant (data not shown). Thus, for the analysis we pooled these two gap sizes together.

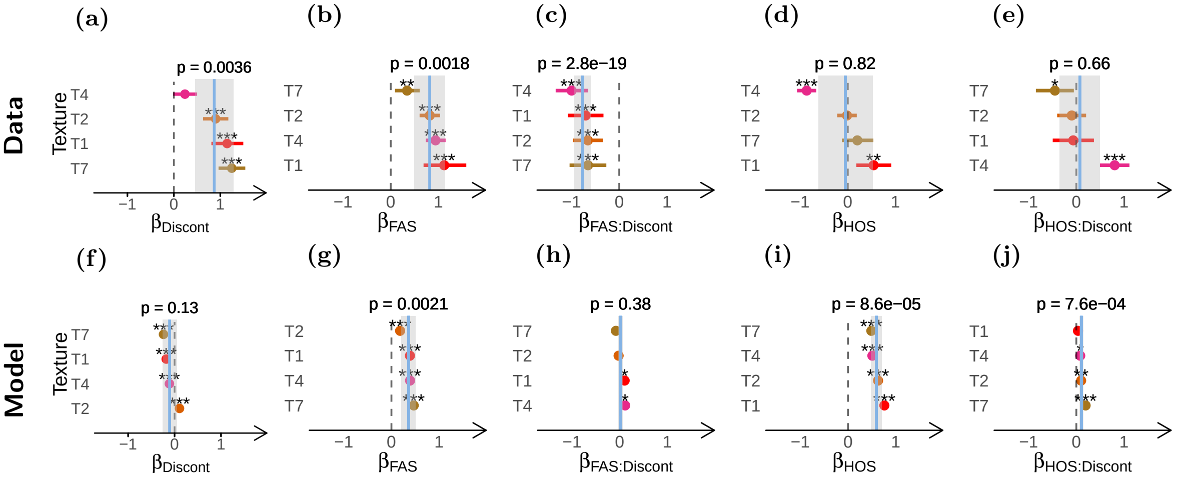


Figure S6: FAS dissimlarity strongly interacts with discontinuity in participants but not in model observers. Parameters estimated by fitting a full model to the data from both continuous and discontinuous conditions in experiment 3. **(a)**-**(e)** LOR for, respectively, discontinuity ($\beta_{Discont}$), FAS dissimilarity (β_{FAS}), discontinuity-FAS interaction ($\beta_{FAS:Discont}$), HOS dissimilarity (β_{HOS}) and discontinuity-HOS interaction ($\beta_{HOS:Discont}$) estimated from the experimental data. **(f)**-**(j)** Same as **(a)**-**(e)** but for the model observers. The data used for fitting the model is the same as that shown in Figure 6. The plots in the figure use the same conventions as Figure 6.

S5 Experiment 4

In the experiment comparing naturalistic and phase scrambled surrounds, we included the discontinuous target-surround conditions for the textures T1 and T2, to determine how segmentation and naturalness interact. We fitted a model to the data with terms for naturalness (β_{Nat}), discontinuity ($\beta_{Discont}$) and their interaction ($\beta_{Nat:Discont}$). In this model β_{Nat} is the effect of naturalness for the continuous condition, $\beta_{Discont}$ is the effect of discontinuity for the phase-scrambled surround, and $\beta_{Nat:Discont}$ represents how the effect of surround naturalness changes for the discontinuous condition (or conversely, how the effect of discontinuity changes when adding surround naturalness). As reported in the main text, there was a strong and significant effect of naturalness for the continuous condition ($\beta_{Nat} = -0.90$, ci $= [-1.57, -0.24]$, p $= 0.03$, Figure S7b). We also observed an effect of discontinuity for the phase-scrambled surround, that did not reach statistical significance for the hierarchical model ($\beta_{Discont} = 0.61$, ci $= [-0.06, 1.30]$, p $= 0.06$, Figure S7c), but did reach significance individually for each texture. Finally, we also observed a moderate and significant interaction between the two terms

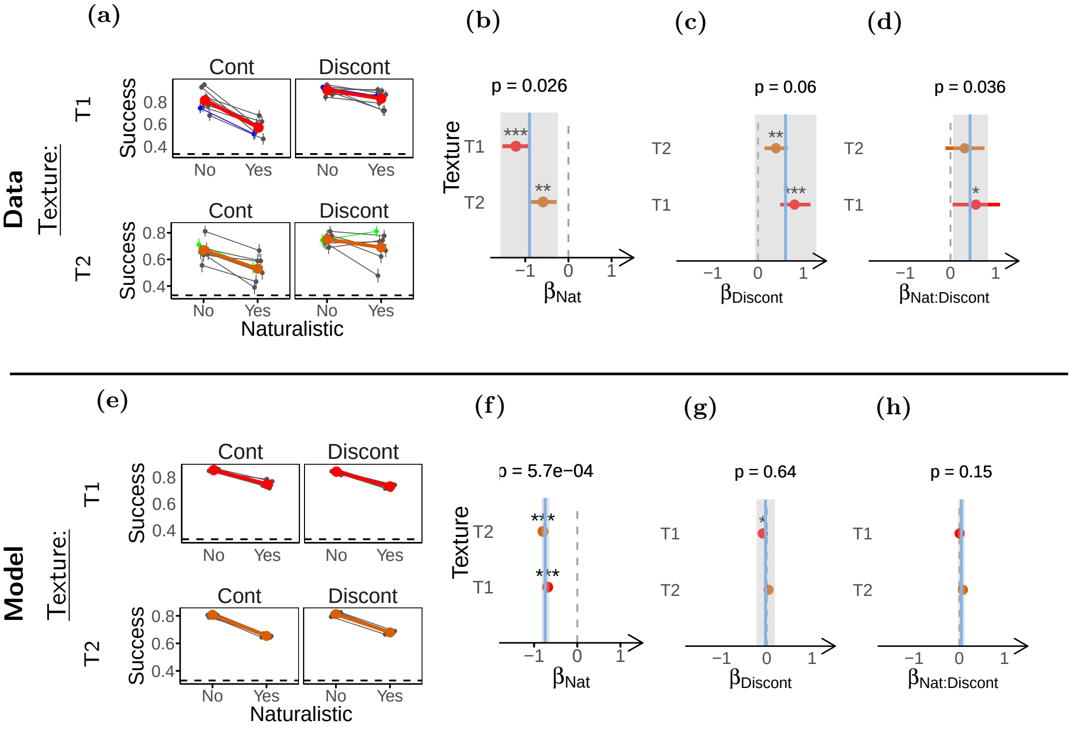


Figure S7: Surround naturalness interacts with segmentation in participants but not in model observers. **(a)** Task performance for the naturalistic and phase-scrambled surrounds in both the continuous (left) and discontinuous (right) conditions. Task performances for textures T1 (top) and texture T2 (bottom) are shown. **(b)-(d)** LOR for, respectively, surround naturalness (β_{Nat}), discontinuity ($\beta_{Discont}$), and their interaction ($\beta_{Nat:Discont}$), estimated from the performance data in **(a)**. **(e)** Same as **(a)** but for the model observers. **(f)-(h)** Same as **(b)-(d)** but for the model observers. Participants (n=28) completed 28 texture-sessions that were included in the analysis, and performed 90 trials per condition.

($\beta_{Nat:Discont} = 0.43$, ci $= [0.06, 0.83]$, p $= 0.04$, Figure S7d).

While the observer model again captured the effect of naturalness ($\beta_{Nat} = -0.74$, ci $= [-0.83, -0.65]$, p $= 6 \times 10^{-4}$, Figure S7f), it failed to capture its interaction with discontinuity, with an interaction close to 0 and non-significant ($\beta_{Nat:Discont} = 0.05$, ci $= [-0.03, 0.12]$, p $= 0.15$, Figure S7h), thus showing that the effect of naturalness is not fully captured by the feedforward version of the SS model.

S6 Texture sampling variability in the model

As described in the methods section, we synthesized large textures with the PS algorithm and then, during the task, patches of these textures were randomly sampled on a trial by trial basis to build the stimuli. This procedure introduces some trial by trial variability in the SS of the displayed patches, and thus this constituted a source of stimulus noise in the task. It is not certain what effect this variability could have on the participants performance, since that would likely depend on how they process the stimuli, and on the strategy they use to solve the task. Nonetheless, we propose there are different reasons to think that this noise does not have an important effect on participants.

First, participants quickly learned from one or two examples how to perform the task, and they could generalize this to the different texture samples. Furthermore, it is very easy to solve the task with foveal inspection, while the task is difficult with limited exposure and peripheral vision. This suggests that the effect of SS variability is probably small as compared to other limitations imposed on the participants by the task design. Lastly, even if the different samples of texture stimuli were always perfectly matched in the PS statistics, it is likely that this will not be the case for the internal representation of SS in participants. This is because different factors such as eye movements, cortical magnification and the use by participants of SS that are not included in the PS model, could still make these different samples unmatched in their internal SS.

On the other hand, given that the implementation of the SS model both lacks the robustness of biological visual systems, and also has few other sources of noise to compete with stimulus variability, the simulated observers may be affected by the sampling variability. Furthermore, understanding the effect of sampling variability on the model can be informative about the behavior of the model, and about possible effects of stimulus noise. Therefore, we performed new simulations removing sampling variability from the task.

In order to remove the sampling variability from the task, we first synthesized for each simulated participant one small PS texture of 128x128 pixels, and its corresponding phase-scrambled texture to be used as targets. Then, to be used for the surrounds, we synthesized one 448x448 texture for each kind of surround texture needed in the experiment to be simulated. All these textures were synthesized as described in the methods, with the exception of their size. Next we used these images to generate one sample of each kind of stimulus (the combination of a kind of surround with a kind of target) using the centermost region of the synthesized textures. This way the crop was not random, and was the same for each stimulus. For example, the naturalistic target was always the exact same patch of texture for the different surrounds, and a given surround was exactly the same for the two kinds of target.

After generating one stimulus of each kind for each simulated subject, we proceeded as in the main simulations to compute the SS and generate pairs of stimuli to discriminate. This results in one kind of stimulus pair for each observer. Finally, we copied the resulting pairs of stimuli to get the needed number of trials to train and test the model (thus there was no variability among samples), and proceeded as described for the main simulations. This way, there was no variability in the stimuli SS in the task (although there was still between trial variability in the noise added to the model).

We note that we used noise with a larger SD than that used in the main simulations. Instead of adding noise with SD equal to that of the SS across the stimuli, we used noise with 10 times that SD. Also, to be able to fit into the synthesized texture for the target, target size was reduced for the first two experiments, using a target size of 100 pixels for experiment 1 and 90 pixels for experiment 2. Also, the plots showing the estimated effects are shown with the same scale as those in the main text, although the effects for this model were considerably smaller. This was done to facilitate the comparison between models and with the experiments.

First, we observed that the performance of the SS model observer was hindered by the presence of the surround in the absence of sampling variability (Figure S8), replicating the results for the original model and the experiments (Figure 3). Then, we repeated experiment 1, by comparing surrounded stimuli with and without a gap. In this case the model had better performance for the discontinuous condition (with gap) than for the continuous (without gap) stimuli (Figure S9), although the effect was very small and non-significant (but it becomes significant when using larger samples). This behavior of the model is the opposite to that observed in the task with stimulus variability (Figure 4), and is opposite from how we expected the SS model to behave, given that the continuous condition has a larger area of informative target texture. Despite its small magnitude, this result underscores the unpredictability of the effect of image manipulations on the SS model, which should be particularly problematic for non-texture stimuli as discussed in the main text.

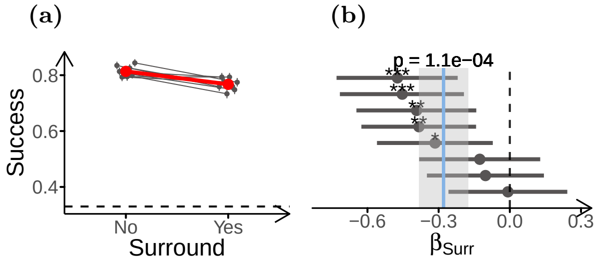


Figure S8: Performance of the SS model is impaired by surround texture in absence of sampling variability. **(a)** Task performance of the SS model for stimuli with and without surrounds, in absence of sampling variability. Same layout as Figure 3b. **(b)** LOR for the presence of the surround. Compare to Figure 3c. 8 model observers were tested on 1500 trials per condition.

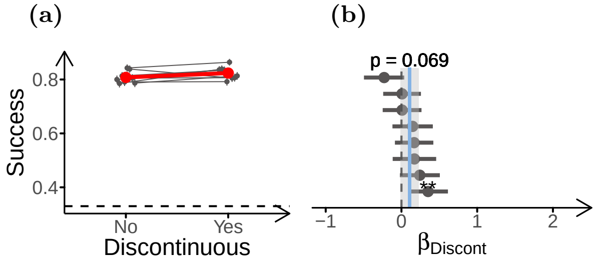


Figure S9: Performance of the SS model is improved by a discontinuity inducing gap in absence of sampling variability. **(a)** Task performance of the SS model for the surrounded stimuli with continuous and discontinuous surrounds. Same layout as Figure 3b. **(b)** LOR for discontinuity ($\beta_{Discont}$). Compare to Figure 4c. 8 model observers were tested on 1500 trials per condition.

Although the previous result could raise doubts about the inability of the SS model to explain segmentation by a gap, when testing the effect of discontinuity by changing target shape (experiment 2) we find that model performance is not affected by discontinuity (Figure S10). This result is in agreement with the original simulations, and different from behavior of the participants, which showed better performance for the discontinuous condition (Figure 5). Thus, overall, we still observe that the SS model cannot capture segmentation by a gap.

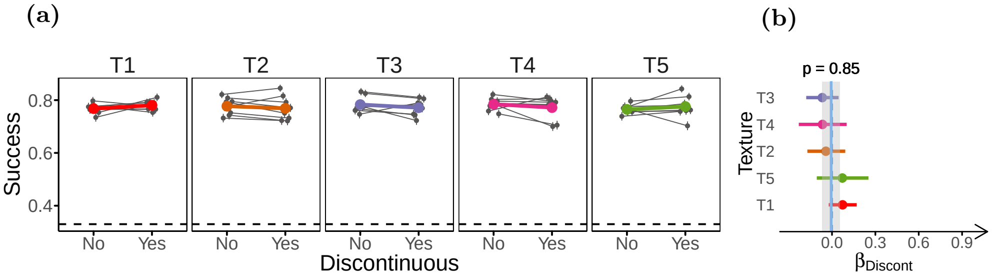


Figure S10: Performance of the SS model is not affected by discontinuity induced by target-shape in absence of sampling variability. **(a)** Task performance of the SS model for the surrounded stimuli with continuous and discontinuous surrounds. Same layout as Figure 3b. **(b)** LOR for discontinuity ($\beta_{Discont}$). Compare to Figure 5c. 8 model observers were tested on 1500 trials per condition.

Next we evaluated the behavior of the model for the different target-surround dissimilarities (experiment 3) in absence of variability. The model did not show effects for FAS dissimilarity, or its interaction with discontinuity (Figures S11b, S11c), while it did show a significan albeit small effect for HOS dissimilarity (Figure S11d). This is not an important change as compared to the model with stimulus variability (Figure S6), and therefore does not affect the previous analysis.

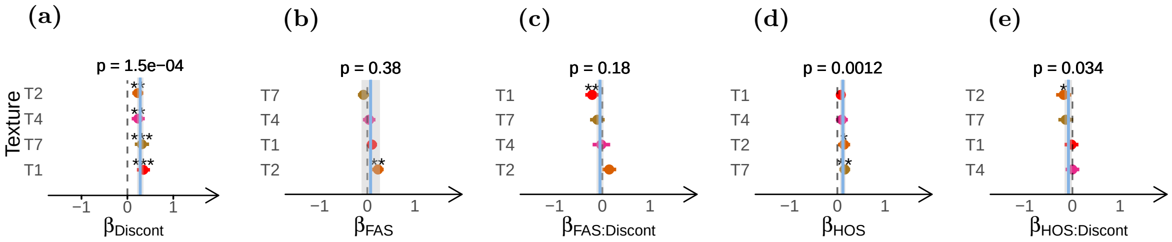


Figure S11: Performance of the model is improved by target-surround dissimilarity in HOS but not in FAS in absence of sampling variability. (a), (b), (c), (d) & (e) LOR for the effects of, respectively, discontinuity ($\beta_{Discont}$), FAS dissimilarity (β_{FAS}), the interaction between FAS and discontinuity ($\beta_{FAS:Discont}$), HOS dissimilarity (β_{HOS}), and the interaction between HOS and discontinuity ($\beta_{HOS:Discont}$). Compare to the estimated LOR shown in Figure S6. Same layout as Figure 5c. 8 model observers were tested on 1500 trials per condition.

Lastly, we tested the model on surround naturalness (experiment 4) without stimulus variability. In this experiment sampling variability could be a particularly important factor for the model, because the sampling of the naturalistic images induced a higher variability in the SS than the phase-scrambled images (analysis not shown). In line with this rationale, we observed that the performance of the SS model was not affected by surround naturalness (Figure S12), unlike what was observed in presence of stimulus variability (Figure 7). Thus, whether the SS model can explain the effect of naturalness that was observed for the participants depends on whether the experimental effect is produced by stimulus variability or not. Given that we argued that stimulus variability due to sampling is probably not an important factor for the participants, this means that the SS model may not explain the effect of naturalness in the experimental data.

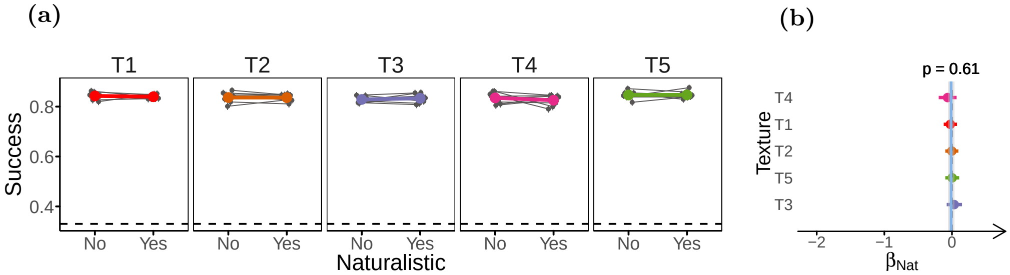


Figure S12: Performance of the model is not affected by naturalistic HOS in the surround, in absence of sampling variability. (a) Task performance of the SS model for surrounds with the presence or absence of naturalistic HOS. (b) LOR for surround naturalness (β_{Nat}) estimated from the data in (a). Compare to Figure 7. 8 model observers were tested on 1500 trials per condition.

S7 Size adjustment results

A difficulty adjustment procedure was carried out for each individual participant before each experiment (except for texture T1 in experiments 1 and 3), to ensure a similar performance level across participants. We adjusted target size using the stochastic approximation procedure [81]. In this procedure, failed responses lead to an increase in target diameter, and successful responses to a decrease, with progressively smaller diameter changes. We set the relation of failure:success step sizes to 9:1, which leads to the procedure converging at a performance of 90% correct responses. The initial target diameter was set between 3.5° and 3.8°, and the initial step size for failed trials was set between 0.7° and 1.0°. The procedure lasted 80 trials or until the step size for the failed trials was smaller than 1 pixel. The actual adaptive procedure started after an initial 10 trials where diameter did not change (totaling 90 trials with the procedure). Figure S13 and Table S1 show the sizes resulting from this adjustment procedure. shows the sizes resulting from this adjustment procedure.

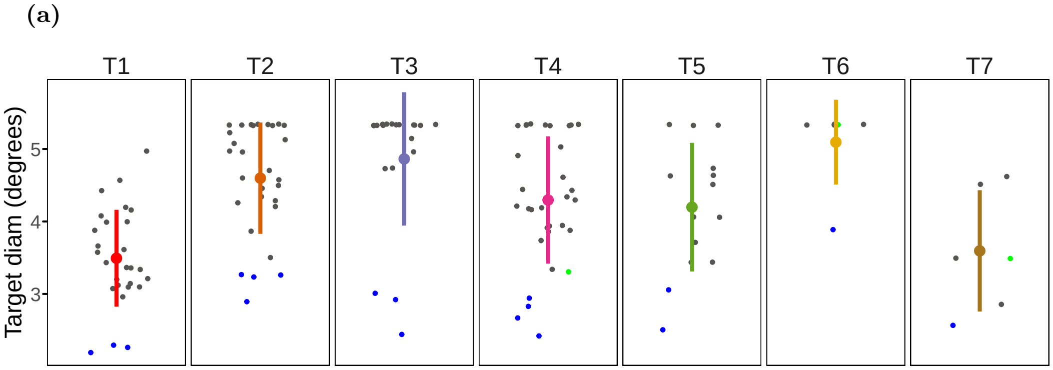


Figure S13: Target sizes obtained after the difficulty adjustment procedure. Smaller dots indicate the target size for a given texture session. Blue dots indicate a texture session for author DH and green dots for author LG. Larger colored dots with vertical lines indicate the mean ±SD accross participants. We excluded from the plot the sizes for texture T1 in experiments 1 and 3, in which size was not adjusted, and for the experiment carried out with only one target.

Texture	Mean (deg)	SD (deg)
T1	3.5	0.67
T2	4.6	0.77
T3	4.9	0.92
T4	4.3	0.88
T5	4.2	0.89
T6	5.1	0.59
T7	3.6	0.84

Table S1: Mean and standard deviation of target sizes obtained by the difficulty adjustment procedure for each texture. The sizes from experiments in which the difficulty adjustment was not performed were not taken into account.

We also tested whether the between participant variability observed in the experimental effects is related to the differences in target size. It is possible that changes in target-surround distances arising from target size adjustment could modify the relative relevance of the contextual modulation processes involved in our task, thus leading to part of the observed between-participant and between-texture variability. We tested this possibility by fitting a linear model to the effect sizes estimated individually for each texture session (that is, for each participant-texture combination, fitting a GLM model only to that texture session data) and the target diameter used in that experimental session. We did this only for experiments 2, 4 and 5 (for the two targets and for the single target experiment, refered to here as 5b), which were the experiments in which only one experimental effect was estimated. For the experiments with multiple textures (all but 5b) we included a term for texture in the linear model, but removing this term does not change the conclusions from the analysis.

Experiment	Correlation	p	Df
2	0.25	<0.01	34
4	0.02	0.61	30
5	-0.01	0.50	31
5b	0.11	0.64	13

Table S2: Pearson correlation coefficients between target diameter and the effect sizes estimated for each texture session in the different experiments.

Table S2 shows the estimated effects of target size on the magnitude of the experimental effect for each experiment. We see that there is a significant effect for experiment 2 but not for the other experiments. The positive sign of the effect in experiment 2 indicates that the larger target size is related to a stronger effect in this experiment. Since most of the experiments did not show an effect, we conclude that target size is not an important factor in explaining between-participant variability.

Redundancy between spectral and higher-order texture statistics for natural image segmentation

Daniel Herrera-Esposito[1]✱, Leonel Gómez-Sena[1], Ruben Coen-Cagli[2]

1 Laboratorio de Neurociencias, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
2 Dept. of Systems and Computational Biology and Dominick P. Purpura Dept. of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA
✱ Corresponding author. E-mail address: dherrera1911@gmail.com

Abstract

Visual texture, defined by local image statistics, provides important information to the human visual system for perceptual segmentation. Second-order or spectral statistics (equivalent to the Fourier power spectrum) are a well-studied segmentation cue. However, the role of higher-order statistics (HOS) in segmentation remains unclear, particularly for natural images. Recent experiments indicate that, in peripheral vision, the HOS of the widely adopted Portilla-Simoncelli texture model are a weak segmentation cue compared to spectral statistics, despite the fact that both are necessary to explain other perceptual phenomena and to support high-quality texture synthesis. Here we test whether this discrepancy reflects a property of natural image statistics. First, we observe that differences in spectral statistics across segments of natural images are redundant with differences in HOS. Second, using linear and nonlinear classifiers, we show that each set of statistics individually affords high performance in natural scenes and texture segmentation tasks, but combining spectral statistics and HOS produces relatively small improvements. Third, we find that HOS improve segmentation for a subset of images, although these images are difficult to identify. We also find that different subsets of HOS improve segmentation to a different extent, in agreement with previous physiological and perceptual work. These results show that the HOS add modestly to spectral statistics for natural image segmentation. We speculate that tuning to natural image statistics under resource constraints could explain the weak contribution of HOS to perceptual segmentation in human peripheral vision.

1) Introduction

Scene segmentation is an essential function of visual processing. Grouping visual features together in a segment and separating different segments in a scene requires multiple processes and sources of information. These include gestalt principles such as proximity, similarity, common fate (Wagemans et al., 2012);

geometrical cues such as symmetry and collinearity (Field et al., 1993; Geisler et al., 2001; Sigman et al., 2001); statistical cues related to texture information (Julesz, 1962; Landy & Bergen, 1991; Z. Li, 2002); binocular disparity cues (Bakin et al., 2000; Zhaoping et al., 2009); detection of edges and boundaries between regions (Ben-Shahar, 2006; Wolfson & Landy, 1998); and shape information derived from object recognition and semantic understanding of scenes (Neri, 2014, 2017), just to name a few.

Here we focus on visual texture, which can be defined by the local statistical properties of an image region (Julesz, 1962; Victor et al., 2017). This is a particularly important and well-studied substrate for image segmentation, as reflected in the vast perceptual (Julesz, 1962; Landy, 2013; Landy & Bergen, 1991; Victor, 1994; Victor et al., 2017) and physiological (Knierim & van Essen, 1992; V. A. Lamme, 1995; V. A. F. Lamme et al., 1999; Nothdurft et al., 2000; Roelfsema, 2006) literature and in successful models of human texture segmentation (Bergen & Landy, 1991; Bhatt et al., 2007; Z. Li, 2002; Malik & Perona, 1990; Victor et al., 2017). Notably, most of this work has been focused on studying second-order statistics (represented in the Fourier power spectrum, henceforth spectral statistics), despite abundant evidence that higher-order statistics (HOS) also strongly influence texture perception (Balas, 2006; Freeman et al., 2013; Freeman & Simoncelli, 2011; Hermundstad et al., 2014; Julesz et al., 1978; Portilla & Simoncelli, 2000; Tesileanu et al., 2020; Victor et al., 2013; Victor & Conte, 1996) and are essential to capture the appearance of natural textures (Balas, 2006; Portilla & Simoncelli, 2000). Studies of HOS cues for texture segmentation have used artificial textures, and relatively low-order statistics (Hermundstad et al., 2014; Julesz et al., 1978; Tesileanu et al., 2020; Tkacik et al., 2010; Victor et al., 2013; Zavitz & Baker, 2014). As a consequence, the relevance of HOS for texture-based segmentation remains uncertain, particularly in the context of natural vision.

Texture processing is especially prominent in peripheral vision, and the most influential theory of peripheral vision relies on summary statistics (SS) of textures (Balas et al., 2009; Freeman et al., 2013; Freeman & Simoncelli, 2011; Rosenholtz, 2016). One important instantiation of the SS theory relies on the statistics defined by the Portilla-Simoncelli (PS) algorithm for texture synthesis (Balas et al., 2009; Freeman & Simoncelli, 2011; Portilla & Simoncelli, 2000), which uses marginal pixel statistics, spectral statistics and a specific set of HOS (detailed below) to synthesize textures with naturalistic appearance. The PS instantiation of the SS model uses a filtering stage analogous to the primary visual cortex (V1) followed by a HOS encoding stage (**Figure 1b**), and captures many aspects of peripheral perception (Balas et al., 2009; Ehinger & Rosenholtz, 2016; Freeman et al., 2013; Freeman & Simoncelli, 2011; Rosenholtz, 2016; Rosenholtz et al., 2012) as well as the selectivity of neurons at higher stages of the visual cortex (Freeman et al., 2013; Okazawa et al., 2015, 2017; Ziemba et al., 2016).

Recent work by us and others (Doerig et al., 2019; Herrera-Esposito et al., 2021; Herzog et al., 2015; Manassi et al., 2012, 2013; Wallis et al., 2019) suggests that perceptual segmentation is an important missing component from the SS model of the visual periphery. In particular, we (Herrera-Esposito et al., 2021) observed that segmentation cues improve performance in a naturalistic texture discrimination task, when target textures are surrounded by distractor textures. Remarkably, however, when we introduced a difference in HOS between target and distractor textures, that difference induced little segmentation, on average, if these regions shared the same spectral statistics (although there is some between texture variability in the estimated effect of HOS). This observation raises the following question: why are these HOS only a weak segmentation cue (relative to spectral statistics) to our peripheral visual systems?

Here we test whether this observation reflects a property of natural images' statistics, which may be exploited by the human peripheral visual system through processes of statistical inference (Hindi Attar et al., 2007), or whether it reflects suboptimal processing. Under the first hypothesis, two scenarios are possible. First (**Figure 1D**, top), the HOS might be an unreliable cue for texture-based segmentation, because differences in local HOS between two regions do not reliably correspond to differences in the segments of the scene. In this case, the visual system would learn to weigh the spectral statistics information more heavily than the HOS information, similar to much previous research in visual (Adams & Mamassian, 2004; Jacobs, 1999; Knill & Saunders, 2003; Saarela & Landy, 2012), auditory (Cazettes et al., 2014; Pavão et al., 2020) and multisensory (Fetsch et al., 2012; Gu et al., 2008) cue combination. A second possibility (**Figure 1D**, middle) is that the HOS may be a reliable cue for segmentation, but highly redundant with the spectral statistics. For example, if it seldom occurs that two different neighboring segments in a natural image have similar spectral statistics but different HOS that allow to segregate them, then these HOS would add little information to the process of texture-based segmentation of natural images. Then, using the spectral statistics but not HOS for peripheral segmentation, could be advantageous considering resource constraints (see Discussion). Lastly (**Figure 1D**, bottom), an alternative hypothesis is that both spectral statistics and HOS are informative about segmentation and independent of each other, in which case the smaller weight placed on HOS by peripheral segmentation processes would reflect a combination of inaccurate encoding of the HOS of PS and suboptimal readout of that information.

To test those possibilities, first we studied how spectral statistics and HOS change across natural textures and natural scenes segments. Next, we trained an observer model to solve a classification task using different combinations of spectral statistics and HOS, in which the goal is to determine whether two image patches belong to the same image segment or not (see **Figure 1**). We used both images of composite natural textures, where we defined the ground-truth segmentation, and images of natural scenes with segmentation maps drawn by humans (Martin et al., 2001). Our results provide the first quantification of the relative power of spectral statistics and HOS of the PS model for texture-based segmentation of natural images.

2) Methods:

2.1) Image and segment selection

For the analysis of texture images we used 638 natural texture images obtained from the Brodatz (Brodatz, 1966), Salzburg Texture Image (*Salzburg Texture Image Database (STex)*, n.d.), and the Lazebnik et al. databases (Lazebnik et al., 2005). We converted the color images to grayscale with the *image* package for octave, by extracting the luminance channel of the YIQ color space. We then normalized the pixel values of each image to have a mean of 0.5 and a standard deviation of 0.2 on a scale between 0 and 1. Next, we cropped 4 non-overlapping square patches of 128 x 128 pixels from the vertices of the image, thus obtaining 4 sample patches per texture (a total of 2552 patches).

For the natural scene analysis we used the 500 natural scene images from the Berkeley segmentation database (BSD) (Martin et al., 2001), and their corresponding segmentation maps labeled by a human (we used the first map available for each image). We converted the color images to grayscale with the same procedure as for textures. The segments analyzed for each image were the central segment of the image (the one containing the central pixel) and all its neighbors. To avoid excessive noise in the computed statistics, we filtered out the images in which the central segment had less than 8192 pixels (equivalent to a 128 x 64 pixels). Furthermore, we also filtered out the neighboring segments with less than 4096 pixels (equivalent to a 64 x 64 pixel patch). After this selection procedure, 416 images and a total of 1696 segments were used.

2.2) Pairing image patches and texture segmentation task

Region-based segmentation consists in the process of determining whether two image regions belong to the same segment or not. We modeled this region-based segmentation task using texture as a substrate by employing a classification task on pairs of image patches.

We generated pairs of image patches that could either belong to the same segment or to different segments (**Figure 1**). For brevity, we refer to these pairs as “matched” and “unmatched” respectively. Then, we computed the statistics of the patches (see details below) and, depending on the analysis, we either computed the angle between the vectors of statistics of the two patches (used in **Figure 2**; see subsection 2.3 for details), or we computed the absolute difference between the patches for each of the PS statistics (used in all other figures and tables). The classification task consists in determining whether the two texture patches belong to the same segment or not.

For the texture images we considered the whole image as one segment, and thus built the matched pairs by pairing two patches from the same texture, and the unmatched pairs by pairing two patches from different textures.

For the natural images in the BSD, the matched pairs were obtained by splitting the center patch, and the neighboring patches with more than 8192 pixels, vertically into two halves at the point that produced the most balanced pixel distribution, and pairing the two halves. The unmatched pairs were obtained by pairing the central patch of an image with a neighboring segment.

2.3) Computing texture statistics

For each image patch we computed the summary statistics of the PS model. These comprise a set of marginal pixel statistics, and a set of statistics over the filter outputs of the steerable pyramid (Portilla & Simoncelli, 2000). We used a bank of filters with 4 orientations and 4 scales, and a neighborhood of 7 pixels for computing spatial correlations. For the cropped texture patches we used the original PS code. For the patches of natural scenes we used a modified version of the Freeman metamer model (Freeman & Simoncelli, 2011). The original code first filters an image with the steerable pyramid, and then computes the weighted average of the pairwise products of filter outputs (equivalent to computing correlations), using a predetermined set of regular weighting windows. Our modification consisted in using an irregular weighting window, given by the segmentation map of the BSD image. In both textures and natural scenes we also modified the code to compute correlations where the original models computed covariances because we observed that correlations afforded better performance in the discrimination task.

We separated the statistics of the PS model into 3 groups, following previous work (Portilla & Simoncelli, 2000; Ziemba et al., 2016): pixel statistics, Fourier power spectrum (spectral statistics), and statistics of higher-order (HOS). The pixel statistics are marginal statistics over the pixel values, including mean, variance, and the skewness and kurtosis at different lowpass versions of the image. The spectral statistics are equivalent to pairwise pixel correlations, which are found in the PS model in the central autocorrelation matrices of the image subsampled at different scales, and in the mean modulus of activation of quadrature pairs of complex filters. The rest of the statistics in the PS model, which are not captured by the marginal pixel statistics or by pairwise pixel correlations, are referred to as HOS. These comprise correlations across space, scale, and orientation between the magnitude of complex bandpass quadrature filters (i.e. the energy of the filters), and local phase statistics (Portilla & Simoncelli, 2000). With the parameters we used for the PS model, we obtained 16 pixel statistics, 137 spectral statistics statistics, and 552 HOS.

2.4) Correlation between spectral statistics and HOS

To analyze the correlation between spectral statistics and HOS differences between image regions, we first z-scored each statistic across the BSD patches to have 0 mean and a standard deviation of 1. Then for each pair of patches we computed the angle between their vectors of spectral statistics and the angle between their vectors of HOS statistics (i.e. we computed the angles in the respective 137 and 552 dimensional spaces for the two kinds of statistics). Then we computed the Pearson correlation between these two.

2.5) Training the linear classifier models

All linear classification models using the absolute differences in statistics were trained by ridge regression, using the glmnet package v4.0-2 (Friedman et al., 2019) in R 3.6.3 (R Core Team, 2018). We used the default settings of the package in which the scaling parameter for the penalization is selected by 10-fold cross-validation on the training set. We used misclassification rate as the criterion for both selecting the penalization parameter and training the model. We also performed a weighting of the pairs of images in the training set so that the overall training weight was the same for the two classes. We also normalized each predictor to have unit variance and zero mean in the training set. We performed this procedure both for the models performing the segmentation task, as for the models performing the identification of pairs of patches with useful HOS.

2.6) Training the segmentation models

For the image segmentation task, we trained a family of linear models to classify the pairs of patches using the absolute difference in each statistic between the patches. The subsets of the PS statistics used in each model are indicated in the text.

For the classification of patches from the natural texture images we first separated the texture images into a training and a testing set, randomly assigning 319 texture images to each. Then for each texture image we generated all the unique combinations of pairs of patches for the matched condition (6 combinations). Then we generated pairs of patches from different textures (randomly sorted) within each image set, generating 10 pairs of these for each texture. This procedure generated 1914 matched pairs of patches and 1595 unmatched pairs of patches for each the training set and the testing set.

For the classification of patches from natural scenes, we randomly sorted the images into a training set of 332 images and a testing set of 84. We then generated the pairs of patches as described above, producing 2688 pairs of patches (1408 matched paris and 1280 unmatched). On average, there were 2150 pairs of patches in the training set, and 537 pairs in the testing set (there is some variability due to the image sorting, since not all images had the same number of segments).

We repeated the random sorting of training and testing set 20 times for each model. In the figures, we show the results for the model trained with each sorting, as well as the average performance.

2.7) Identifying pairs with useful HOS

To identify the pairs of patches where HOS improved segmentation (referred to as pairs with useful HOS for brevity), we first split the number of images in the dataset (either for BSD or for the textures) into 10 non-overlapping subsets, to be used as testing sets separately. Then, we iterated through all the 10 subsets of patches, training a segmentation model on the image pairs that did not belong to the testing subset, and then testing the model on the subset. For each subset we trained both a model using spectral statistics alone, and a model using HOS alone. Then, for each pair in the testing set, we compared the outputs of the two models, and we labeled all pairs of patches that were incorrectly classified by spectral statistics but correctly classified by HOS as having useful HOS. We repeated this procedure for all testing sets, obtaining a label for each pair of patches. The same procedure was performed comparing the model with spectral statistics alone to the model with both spectral and HOS.

Note that the size of the train and test sets used here are different from the main segmentation task. As described above, for the main segmentation task, when using the texture dataset half of the textures went into the training set, and when using natural images, 20% of the images went into the training set. Here, in both cases 90% of the dataset went into training for each model. This means that for the texture dataset, 6314 pairs of textures were used for training, and 704 for testing in each iteration. For the natural scenes dataset, on average 2419 pairs went into the training set and 269 into the testing set for each iteration.

Then, we again split the dataset into 10 subsets, and for each subset, we trained a linear classifier on the rest of the patches to identify whether the pairs had useful HOS or not, using as inputs for this task the spectral and HOS. This way, for each pair of patches we obtained a ground-truth label indicating whether it had useful HOS, as observed in the segmentation models, and the output of a classifier labeling it as having useful HOS or not.

2.8) Data and code availability

All the analysis code used in this work is available at https://git.io/JJNyr.

3) Results:

We used a texture discrimination task to quantify the contribution of different image statistics to texture-based segmentation (**Figure 1**). Specifically, the texture statistics of two image patches are given as input, and a classifier indicates whether these two patches belong to the same image segment or not (matched and unmatched pairs of patches respectively). We considered different groups of image statistics of the PS texture model: marginal pixel statistics, Fourier power spectrum (spectral statistics), and higher-order statistics (HOS) (see Methods for further detail). To quantify the contribution of these statistics, we trained different models using the absolute difference between the values of these statistics and compared their performance at the task.

3.1) Differences in spectral statistics and HOS are redundant in natural images

We first studied the correlation between differences in spectral statistics and HOS across segments, as a basic estimate of redundancy **Figure 2** shows, for 2688 pairs of neighboring image patches sampled from 416 natural scenes (BSD, (Martin et al., 2001), the angle between their vectors of spectral statistics (which measures how different the spectral statistics are between the two patches), and the angle between their HOS statistics. We found a strong positive correlation between the spectral statistics angles and the HOS angles (**Figure 2**; Pearson correlation = 0.55, p = 2e-16, CI = [0.52-0.58]) for neighboring patches, suggesting a high redundancy between these statistics.

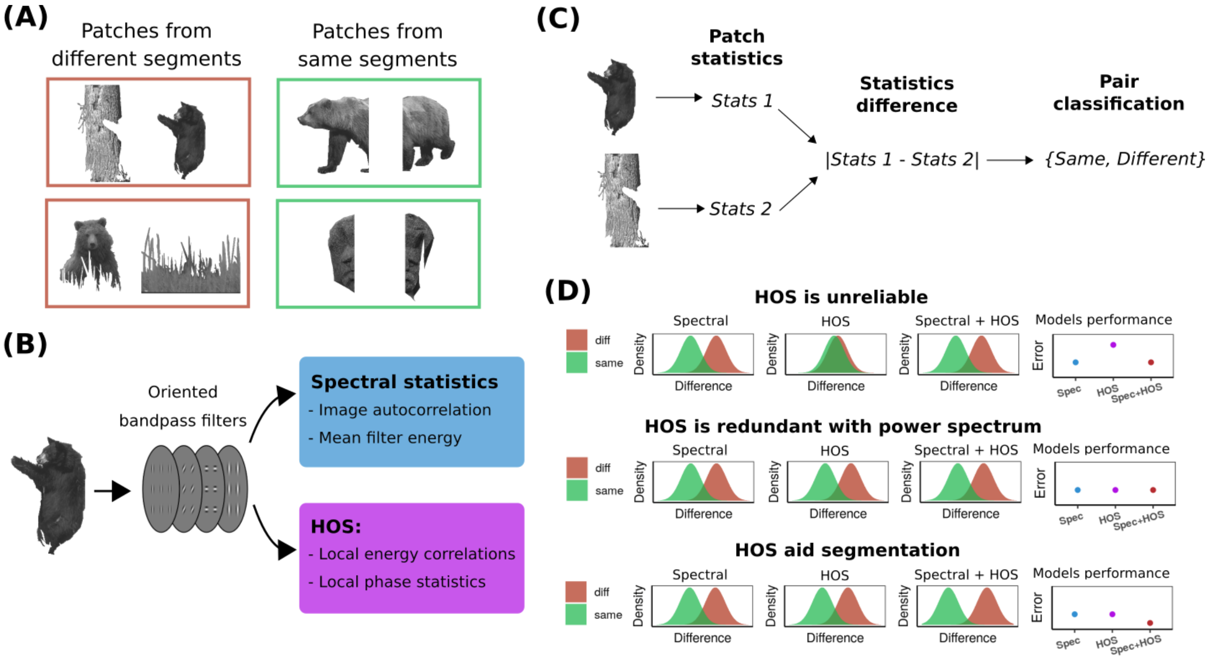


Figure 1. (A) Example pairs of patches taken from different (left) or from the same (right) image segment of a natural image. (B) Illustration of the computing of the different image statistics, with a first filtering stage and a second stage of computing image statistics. (C) Segmentation task, in which the statistics of two image patches are used to classify them as belonging to the same or to different segments. (D) Each row illustrates one possible scenario of spectral statistics and HOS contributions to image segmentation. Plots show the distribution of the difference between statistics across image patches from the same (green) and from different (brown) segments for different combinations of statistics (first three columns), and the corresponding performance of different segmentation models using these statistics (fourth column).

Besides the correlation, which indicates overall redundancy, a more relevant question is how much information the differences in the individual spectral statistics and HOS provide for the task of segmentation. To quantify this we next measured how the use of the spectral statistics difference compares to both the use of HOS, and to the combination of spectral statistics and HOS for segmentation.

3.2) Spectral statistics and HOS are redundant for natural scene segmentation

To test the information in the different sets of statistics and in their combination for image segmentation, we trained a linear classifier on the individual statistics of the PS texture model using ridge regression to solve a segmentation task (see Methods). We used the same 2688 pairs of natural scene patches as in **Figure 2**. We performed 20 repetitions of the task by randomly separating the image set into 332 images for the training set (an average of 2150 pairs of segments) and 84 for the testing set (an average of 537 segments) for each repetition.

Figure 3B shows that adding the spectral statistics to the marginal pixel statistics improved performance, albeit modestly. The combination of pixel and HOS performed better than pixel and spectral statistics, although the difference was small, and spectral statistics performed slightly better than HOS when both were used without pixel statistics. Finally, combining spectral and HOS led to an improvement in segmentation performance (both with and without pixel statistics), although the improvement was modest.

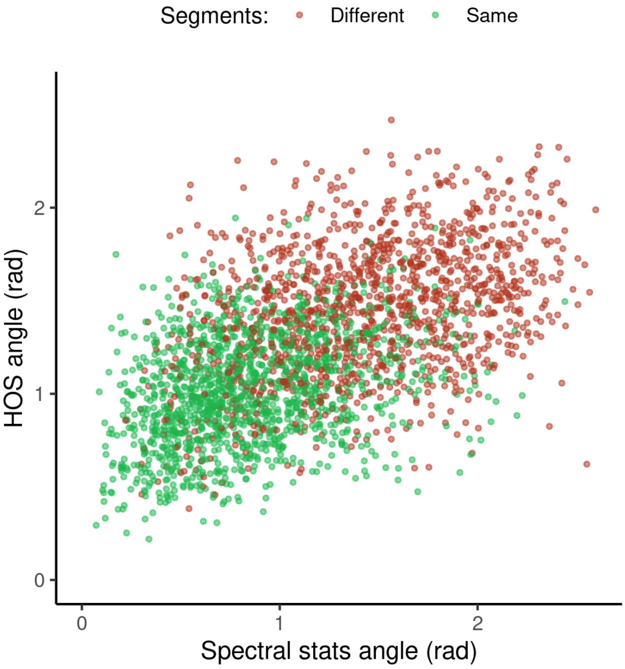


Figure 2. Relation between the difference in HOS and the difference in spectral statistics for pairs of image patches from the BSD database. The color of the dots indicates whether the pair of patches are extracted from the same image segment (green) or not (brown).

These results show that the segmentation performance of the model using spectral statistics is high, and that it improves only modestly when adding HOS, even though HOS alone also achieved high performance. This supports the idea that the HOS of the PS model are highly redundant with the spectral statistics for segmenting natural images. We observed similar results with a non-linear decoder (i.e. a neural network), showing that our findings do not simply reflect a limitation of the linear decoder (**Tables S1, S2**, **Figure S1**).

We reasoned that our results could be influenced by differences in dimensionality: the spectral statistics of the Portilla-Simoncelli model are 4 times less numerous than the HOS (137 statistics and 552 statistics respectively, with our selected number of orientations, scales and neighborhood size), and we found similar ratios for their intrinsic dimensionality (**Table S3, S4**). To test this possibility, we used PCA to match the dimensionality of the spectral and HOS statistics, and we found similar results to **Figure 3B** (**Table S5**). Thus, despite the HOS being more numerous, when using subspaces with the same dimensionality, spectral and HOS statistics still perform similarly on segmentation.

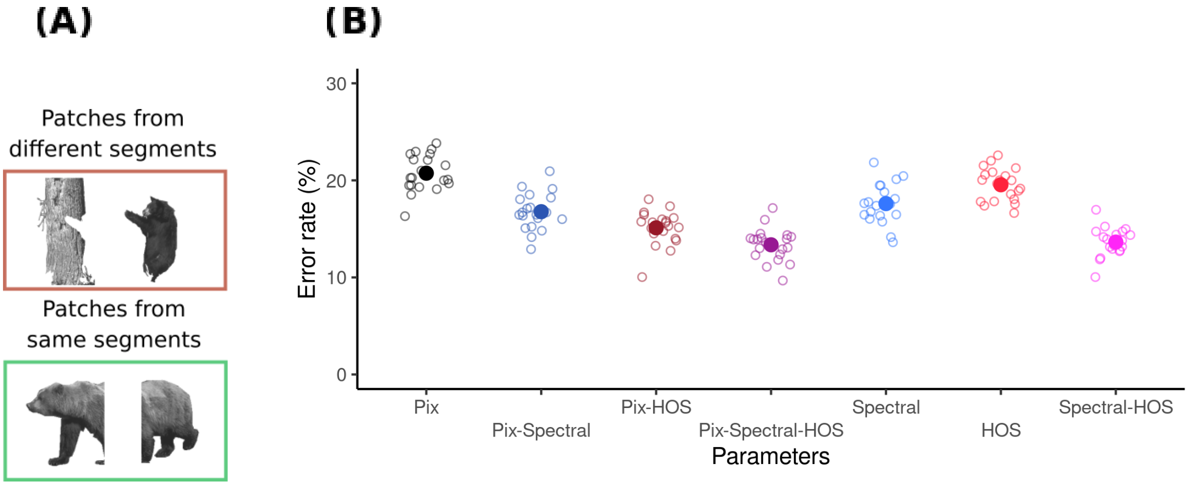


Figure 3. (A) Example of pairs of image patches used in the task. Top: Patches belong to the different segments. Bottom: Patches belong to the same segment. (B) Performance of a linear model in classifying pairs of image patches as belonging to the same or to different image segments. Models using different subsets of statistics from the PS model are shown. The empty circles show model performance for the 20 individual models trained and evaluated with different splits of training and testing data sets. The filled circles show the mean performance across splits.

3.3) Spectral statistics and HOS are redundant for natural texture segmentation

Next, we asked whether the contributions of these sets of statistics to the more specific task of segmenting natural textures, is similar to what we found for natural scenes. This is important because in the BSD, natural scenes have been segmented manually by human observers who likely used several other cues, in addition to texture, to determine segmentation. The use of these other cues for segmenting and grouping image patches in scenes may influence the observed distribution of texture features across and within segments. Thus, to better understand the contribution of HOS to natural texture segmentation, we next repeated the analysis for a texture segmentation task, using pairs of patches that were obtained from natural texture images (Brodatz, 1966; Lazebnik et al., 2005; *Salzburg Texture Image Database (STex)*, n.d.) (**Figure 4A**).

We trained the model on 3509 pairs of texture patches using ridge regression (see Methods), and then tested it on 3509 pairs of patches sampled from a different set of natural textures. We repeated this procedure 20 times, resampling the training and test sets.

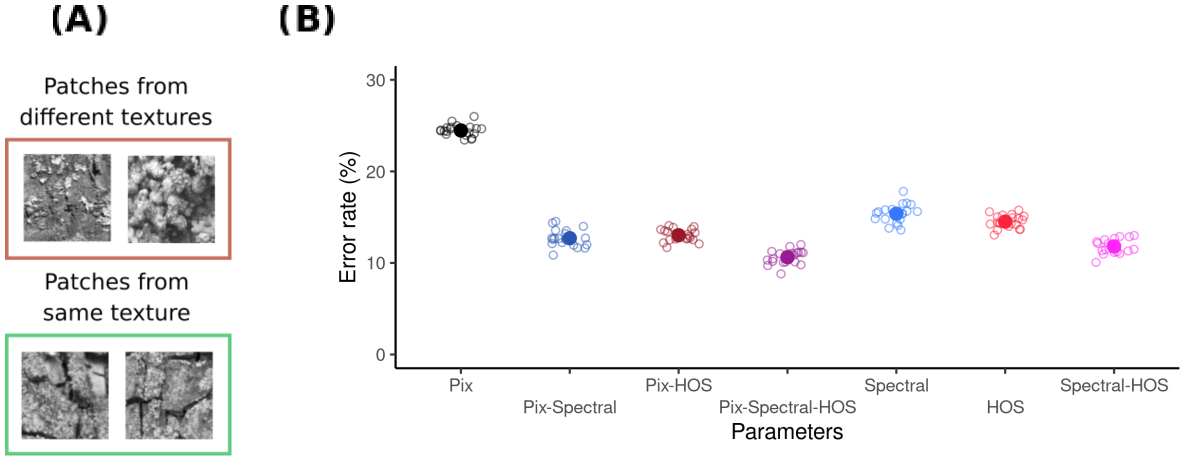


Figure 4. (A) Example of pairs of texture patches used in the task. Top: Patches belong to the different textures. Bottom: Patches belong to the same texture. (B) Performance of a linear model classifying pairs of texture patches as belonging to the same or to different textures. Models using different subsets of the PS statistics are shown. Same conventions as in Fig. 3B.

We observed that segmentation performance for natural textures was in general higher than for the natural scenes (**Figure 4B**). Also, segmentation improved substantially when we added the spectral statistics to the pixel statistics. As with natural images, we observed that using pixel statistics and HOS led to similar performance than using pixel and spectral statistics. Using both the spectral statistics and HOS also improved performance over using the spectral statistics alone, although modestly.

These results indicate that spectral and HOS are redundant for the task of natural texture-segmentation. We note that although the results for texture segmentation are similar to those for scene segmentation, models trained on one dataset generalize poorly to the other (**Table S6**), supporting the hypothesis that different texture features may be required for the two tasks.

3.4) Images with useful HOS are difficult to identify

In our experimental work (Herrera-Esposito et al., 2021), we observed considerable between-texture variability in the effect of HOS on segmentation. Similarly, previous work (Freeman et al., 2013; Okazawa et al., 2015, 2017; Ziemba et al., 2016) showed that different synthetic PS textures lead to different perceptual and neural discriminability. Therefore, we considered the possibility that, although combining spectral and HOS leads to a modest performance improvement overall, HOS could be particularly useful for some subset of images.

To analyze this possibility, we identified pairs of natural scene patches where classification was better when using HOS. Specifically, we compared for each pair of patches the classification with spectral statistics alone to HOS alone (we obtained similar results when using spectral and HOS together, data not shown).

	Label: No HOS improvement	**Label: HOS improvement**	
Predicted: No HOS improvement	1707	160	1867
Predicted: HOS improvement	711	110	821
	2418	270	

Table 1. Classification of pairs of natural image patches as being better segmented by HOS or not. The columns of the table indicate the observed ground truth label, of whether a pair of patches was better labeled by HOS than by spectral statistics or not. The rows indicate the label for the pairs of patches predicted by a linear classifier. Each cell in the table shows the number of pairs for each combination of true and predicted labels.

Overall, 10% of the full set of pairs were better classified by HOS than by spectral statistics, with similar proportions for pairs belonging to the same and to different segments (data not shown). Conversely, 12.5% of the pairs were better classified by spectral statistics than by HOS. In particular, 59% of the pairs misclassified by spectral statistics were correctly classified by HOS, confirming that HOS are useful for some images. In addition, if segmentation by spectral and HOS were independent, the error rates for HOS should be the same in the complete dataset (19.3%) as in the subset misclassified by spectral statistics. Instead, 41% of the pairs misclassified by spectral statistics were also misclassified by HOS, reflecting the redundancy in the responses of the two sets of statistics.

We next tested whether this subset of images with more useful HOS can be identified from their statistics, which would be required for the visual system to use HOS more strongly in these cases. For this, we relabeled the pairs of patches to indicate whether segmentation was better when using HOS or not, and we then trained a new linear classifier on these labels, using spectral and HOS together as predictors (see Methods).

The confusion matrix (**Table 1**) shows that the classifier performed better than chance, ($p < 2e-16$, McNemar’s test), which indicates that there is some consistent difference between pairs where HOS improve segmentation and those where it does not. However, due to the imbalance between the classes, we observe a low overall accuracy of 68%, that is lower than the accuracy obtained by classifying all pairs as not being improved by HOS. In line with these results, visual inspection of the pairs of patches better classified by HOS does not show obvious patterns that distinguish them from other pairs (**Figure 5A**).

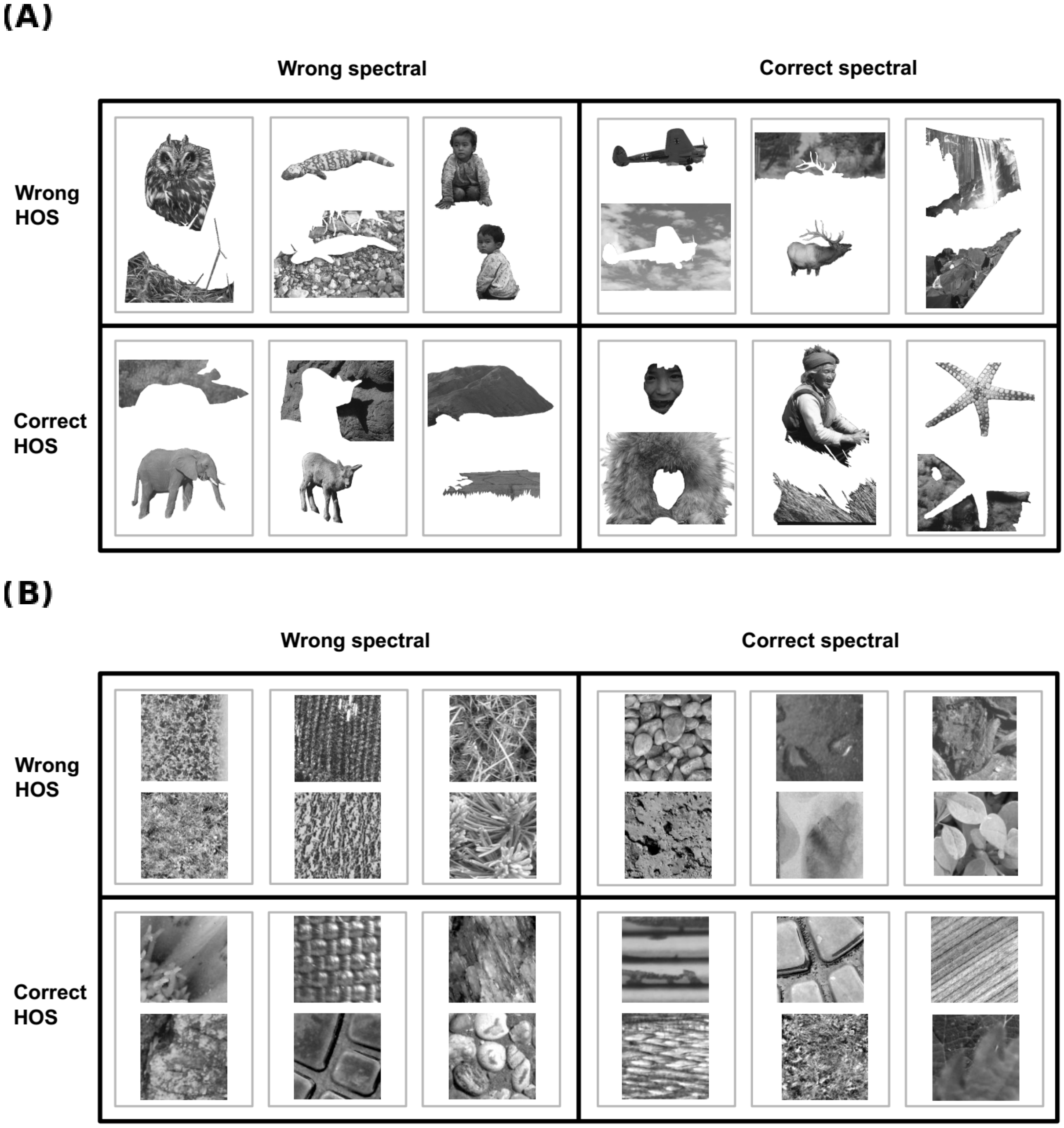


Figure 5. (A) Examples of pairs of natural scene patches with different classification outcomes for both spectral and HOS. Only pairs extracted from different segments are shown. Smaller gray boxes group the patches that form a pair. Larger black boxes indicate the classification outcome for both spectral and HOS. **(B)** Same as **(A)** but for natural textures.

We obtained similar results for texture segmentation. We found that 8.7% of the texture pairs were better classified by HOS than by spectral statistics, and that 8.4% were better classified by spectral statistics than by HOS. Of the pairs misclassified by spectral statistics, 58% of which were also misclassified by HOS, showing again redundancy in their responses. A classifier trained to identify the patches with HOS improvement, as described for BSD above, had a performance better than chance ($p < 2e-16$, McNemar's test, **Table 2**), but with a low accuracy of 67%. **Figure 5B** shows some example pairs of textures with different classification outcomes for spectral and HOS (more example pairs can be found together in the open repository with the analysis code).

	Label: No HOS improvement	**Label: HOS improvement**	
Predicted: No HOS improvement	4324	271	4595
Predicted: HOS improvement	2087	336	2423
	6411	607	

Table 2. Classification of pairs of texture image patches as being better segmented by HOS or not. Same conventions as **Table 1**

In sum, the misclassifications of spectral and HOS showed redundancy in both natural scenes and textures, but a subset of the pairs of patches was better classified by HOS. Nonetheless, the pairs better classified by HOS were not accurately identified by a linear classifier (for further analysis on the causes of the low accuracy see Supplementary section **S4**). We also obtained similar results when using a procedure to reduce possible labeling noise, in which we trained to separate models on non-overlapping subsets of the training set, and required that HOS be better than spectral statistics in both training sets for a given pair, in order to label that pair as having useful HOS (results not shown).

Furthermore, we compared the predictions from these models to our previous experimental results (Herrera-Esposito et al., 2021), to test whether the observed experimental variability between pairs of textures correlates with the estimated usefulness of the HOS. We did not observe any clear agreement between the two that could be suggestive of fine-tuning to use HOS in informative cases (**Figure S2**, although the low number of textures and several other caveats demand caution when interpreting these results, see the Supplementary section **S5**).

3.5) Subsets of HOS contribute differently to segmentation

Besides the results from previous studies mentioned above, showing that different PS textures drive mid-level visual areas and perception to different degrees, the same line of research has also identified specific subsets of HOS as driving perception (Freeman et al., 2013; Hermundstad et al., 2014; Tesileanu et al., 2020; Victor et al., 2013) and physiology (Freeman et al., 2013; Okazawa et al., 2015, 2017; Yu et al., 2015) to different degrees. Also, this has been shown to follow natural image statistics (Hermundstad et al., 2014; Tesileanu et al., 2020; Yu et al., 2015). Therefore, we next wondered whether different subsets of HOS would show varying degrees of usefulness in our segmentation task.

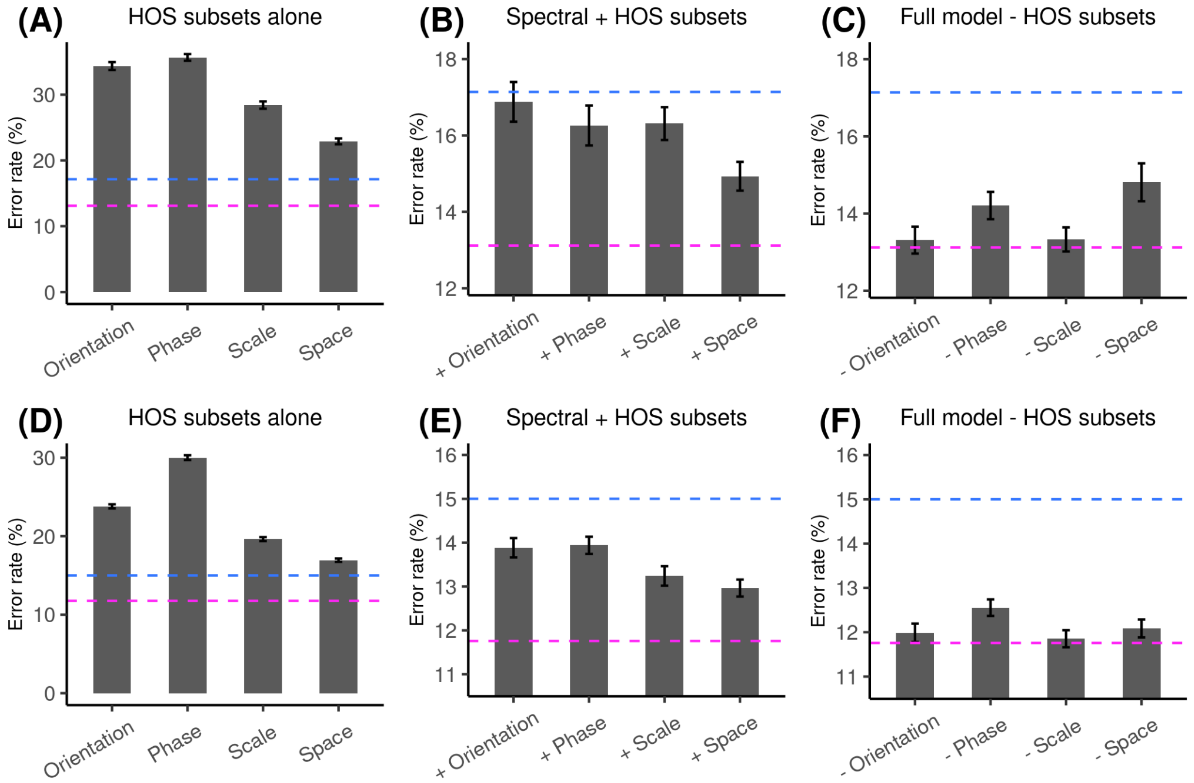


Figure 6. Performance at the segmentation task using different subsets of HOS. Each grey bar shows the average over the outcome of 50 different models trained on random splittings of the data into train and test set. The error bars show the 95% confidence interval of the mean. The dashed horizontal blue line shows the performance for the model using only spectral statistics, and the red line shows the performance of the model with spectral and all HOS. **(A)** Segmentation performance for the BSD using subsets of HOS alone. **(B)** Segmentation performance for the BSD using subsets of HOS together with spectral statistics. **(C)** Segmentation performance for the BSD using the model containing spectral statistics and all subsets of HOS except those indicated in the horizontal axis. **(D)**, **(E)** and **(F)**, same as **(A)**, **(B)** and **(C)** but for the texture segmentation dataset.

To determine the relevance of different subsets of HOS to our segmentation model, we divided the HOS into the four subsets used in the PS model (Portilla & Simoncelli, 2000): energy correlations across space, energy correlations across scale, energy correlations across orientation, and phase correlations across scale (also called linear correlations across scale in some studies). We then tested the performance of different combinations of these subsets of HOS, with and without spectral statistics, in the segmentation task.

Figure 6 shows the performance of the segmentation models (top row, natural scenes; bottom row, natural textures) using different subsets of HOS. All the subsets of HOS alone had considerably worse performance than spectral statistics (indicated by the dashed blue line, **Figure 6A**) for natural scene segmentation. Adding each subset of HOS to spectral statistics did not reach the performance of the full model (purple line, **Figure 6B**). Similarly removing each subset of HOS never decreased performance to the level of spectral statistics alone (**Figure 6C**). In most cases, spatial correlations were the most useful subset of HOS. Also, orientation and phase statistics were the least useful when considered alone and together with spectral statistics, but phase statistics gained in relevance when removing the subsets of HOS from the full model.

Results for textures (**Figure 6D-F**) were similar to those for natural scenes, except that adding orientation correlations improved performance more markedly, and removing spatial correlations did not reduce performance.

These analyses confirm that different HOS subsets have different usefulness for segmenting natural scenes and natural textures. Spatial correlations seem to be, in general, the most informative subset of HOS, and in most cases they were followed by correlations across scale. Phase correlations allowed for improved performance when combined with spectral statistics, and they had a considerable effect when removed from the full model, indicating that they contain useful information that is not redundant with the rest of the HOS. Correlations across orientation were generally among the least useful for segmentation when considering models containing spectral statistics.

4) Discussion

We have studied the importance of different image statistics, namely the spectral statistics and HOS of the PS texture model, for segmenting natural textures and images. First, we showed that there is a strong correlation between the difference in spectral statistics and the difference in HOS for pairs of neighboring patches in natural scenes (**Figure 2**). Then, using segmentation tasks with either natural scenes segmented by human observers or natural textures, we showed that using either the spectral statistics alone or the HOS alone were enough to solve the task with high accuracy, indicating they are both reliable cues for segmentation. Importantly, combining both together produced modest improvements, for both linear and non-linear classifiers (**Figures 2, 3, S1, Table S2**). These results indicate a strong redundancy between spectral statistics and HOS specifically in the context of image segmentation, and seem to rule out the alternatives that HOS cues for segmentation are either unreliable or largely independent from spectral statistics.

In a recent study on human texture perception, we reported that differences in the HOS of the PS model between adjacent textures in peripheral vision produced only weak segmentation when the textures had matched spectral statistics (Herrera-Esposito et al., 2021). In another related study (Balas, 2008), the author observed that human texture similarity judgements were better predicted by the power spectrum of the textures alone, than by the entire set of PS statistics. These results are of particular interest because these statistics have high perceptual relevance (Balas, 2006; Freeman et al., 2013; Freeman & Simoncelli, 2011; Portilla & Simoncelli, 2000; Wallis et al., 2017) and drive neural activity in mid-level visual areas (Freeman et al., 2013; Okazawa et al., 2015, 2017; Ziemba et al., 2016). Furthermore, these statistics are related to the second processing stage in the SS model of peripheral vision (Balas et al., 2009; Freeman et al., 2013; Freeman & Simoncelli, 2011; Rosenholtz, 2016), of which segmentation has been argued to be an important missing component (Doerig et al., 2019; Herrera-Esposito et al., 2021; Herzog et al., 2015; Manassi et al., 2012, 2013; Wallis et al., 2019), making their role in segmentation an essential aspect for the further development of this model. In the present work we expand on our previous results showing that the small effect observed for these HOS on perceptual segmentation may be related to their redundancy with spectral statistics in natural images for the task of image segmentation (**Figure 1**), since they may not add much to the initial segmentation process based on the power-spectrum representation in V1 (V. A. Lamme, 1995; Landy & Bergen, 1991; Z. Li, 2002; Nothdurft et al., 2000; Victor et al., 2017).

In line with this argument, previous work showed that the higher variability in second-order pixel statistics in natural images as compared to third and fourth-order pixel statistics matched their perceptual saliency (Hermundstad et al., 2014; Tesileanu et al., 2020; Tkacik et al., 2010). Nonetheless, besides using a different kind of texture than the ones presented in this work and our experimental study (Herrera-Esposito et al., 2021), the computational analysis of image statistics in these previous studies was performed in the context of efficient coding, rather than the specific perceptual task of image segmentation. Different tasks may rely on different texture properties (Victor et al., 2017), which can explain why the HOS of the PS model are simultaneously very important for texture perception (Balas, 2006; Portilla & Simoncelli, 2000; Wallis et al., 2017) but maybe less so for segmentation. A variation of this idea is also espoused in those previous studies on natural texture statistics (Hermundstad et al., 2014; Yu et al., 2015), where it is noted that the sensory periphery (i.e. the retina) and the cortex face different constraints and goals that lead to different coding regimes. In this sense, the present work is a contribution to the growing efforts of comparing perceptual systems to model observers performing sophisticated tasks on natural images (Burge, 2020).

But although there is a high redundancy between the spectral statistics and the HOS for image segmentation, the observation that HOS are a reliable segmentation cue and that they can improve texture segmentation, raises the question of why the spectral statistics and not the HOS are used as a strong segmentation cue, as shown in peripheral vision (Hermundstad et al., 2014; Herrera-Esposito et al., 2021; Victor et al., 2013). One possible explanation to this regard is the constraint in resources that makes information processing by the visual system a balance of costs and benefits. While the HOS improved model performance, they did so only modestly and one could hypothesize that the cost of using these HOS would lead the visual system to use the spectral statistics as the main segmentation cue, with these HOS being a secondary or null segmentation cue.

Nonetheless, a softer version of the hypothesis is that the HOS of the PS model are particularly useful in some scenarios (i.e. specific kinds of images), and that the visual system is fine-tuned to rely on HOS in these cases. As mentioned previously, this could be in line with our previous experimental work on perceptual human segmentation (Herrera-Esposito et al., 2021), as well as with previous physiological and perceptual work studying PS textures (Freeman et al., 2013; Okazawa et al., 2015, 2017; Ziemba et al., 2016). Since this fine-tuning would rely on the ability to identify which images have useful HOS for segmentation, here we tested whether a linear model could identify the pairs of patches where HOS improved segmentation over spectral statistics. We found that these pairs of patches were difficult to identify (**Tables 1, 2**), suggesting that this kind of fine-tuning may be difficult to achieve in practice. Furthermore, when comparing the predictions from the models to our experimental results reported in (Herrera-Esposito et al., 2021) we did not find any agreement between models and experiment that could suggest such a fine tuning (**Figure S2**, although this analysis is preliminary due to the little experimental data available, and should be interpreted with caution).

Although the pairs of patches with useful HOS could not be clearly identified, we did find that some subsets of HOS are more useful than others for segmentation, both in isolation and in combination with spectral statistics (**Figure 6**). Mainly, we observed that when considering the subsets of HOS separately, spatial and scale energy correlations were generally the ones with best segmentation performance (**Figures 6A, 6B, 6D, 6E**). This finding agrees with previous studies showing that scale and spatial energy correlations explain the most variance in the variability of perceptual sensitivity between PS textures (Freeman et al., 2013), and in V4 neurons ability to discriminate PS textures from noise (Okazawa et al., 2015).

The agreement between our analysis and these previous results may reflect a fine tuning of the visual system to the usefulness of the different subsets of HOS, which is captured in our analysis of segmentation. But this agreement does not necessarily mean that the visual system is tuned to use these HOS specifically for segmentation. One alternative is that these HOS are the most informative ones in general, and the visual system is tuned to use them for other tasks as well. In relation to this, (Okazawa et al., 2015) reported that energy correlations across space are the HOS with highest performance in a texture classification task, which they propose as a possible explanation to their physiology results. Furthermore, the ordering of HOS relevance may also depend on what ranking criterion is used, requiring careful comparisons across tasks and methods. For example, we observed a different ordering of the relevance of HOS subsets when performing segmentation alone than when in the context of the full model (**Figures 6C, 6F**). In line with this, the ranking of HOS subsets relevance obtained from analyzing their contribution to discriminability of textures from spectrally matched noise in V4 is somewhat different from the ranking obtained for explaining V4 responses to textures in general (Okazawa et al., 2015, 2017). Therefore, more work is needed to understand how the information different HOS subsets carry for segmentation in natural images, relates to their use by the visual system.

In conclusion, the results presented here, show that spectral statistics and the HOS of the PS model have a strong redundancy for natural scene and texture segmentation, which coupled with resource constraints may explain the weak effect of these HOS in human segmentation (Herrera-Esposito et al., 2021). This also suggests that segmentation based on the HOS of the PS model may not be crucial to future extensions of the SS model of peripheral vision, but rather that existing models of segmentation based on the outputs of V1-like oriented filters that respond to spectral statistics (Bergen & Landy, 1991; Bhatt et al., 2007; Z. Li, 2002) may be enough to considerably expand its explanatory power. Nonetheless, there are some important caveats that need to be considered.

One important caveat is that the redundancy between spectral and HOS reported here is compatible with either of them taking a secondary role. Although there is plenty of evidence showing the primacy of spectral statistics over HOS in texture segmentation, these studies have been mostly done in the peripheral visual field. Therefore, it is not clear whether the same holds for central vision, and our results here do not necessarily mean that HOS take a secondary role there. Furthermore, our main line of argument rests on the potential cost of using HOS for segmentation, and the role of resource constraints in the brain. Given that resources are much more constrained in the periphery than in central vision, our line of thought is compatible with a stronger role of HOS for segmentation in central vision.

Another important caveat is that we only considered a specific set of HOS, the ones in the PS model. While the HOS of the PS model capture to a considerable extent the perceptual quality of natural textures, they sometimes fail to fully reproduce their structure (Portilla & Simoncelli, 2000). Therefore, other HOS not present in the PS model are important for texture perception, and it is possible that they contribute more strongly to segmentation, both in humans and in segmentation models. One example is the correlations between the features of mid-level layers in deep neural networks, which have been shown to capture the visual appearance of many textures (Gatys et al., 2015), and that allow for good performance in image segmentation (Vacher & Coen-Cagli, 2019).

On the other hand, we also did not consider other low-level segmentation (or saliency) cues that are represented in V1, such as color, binocular disparity and motion (Braddick, 1993; A. Li & Lennie, 2001; Møller & Hurlbert, 1996; Nakayama et al., 1989; Saarela & Landy, 2012). A more general version of our main hypothesis could be that, for segmentation, the HOS of the PS model are redundant with the cues available in V1 in general, and not only with spectral statistics. This alternative hypothesis would be more in line with the proposal that there is a bottom-up saliency or segmentation map in V1 based on these features (Z. Li, 2002; Zhang et al., 2012; Zhaoping, 2019). Therefore, it is possible that by ignoring these other early segmentation cues, we overestimated the contribution of HOS to bottom-up segmentation. This more general hypothesis may also explain why, being redundancy a mutual relationship where either kind of statistics could be used, it is the HOS that adopt a secondary role.

The last important consideration is that we used a region-based texture segmentation task (i.e. using the properties of two image regions to decide whether they belong to the same segment), but segmentation may also proceed through processes based on identifying texture-defined edges (Giora & Casco, 2007; Landy, 2013; Machilsen & Wagemans, 2011; Rosenholtz, 2014). This other type of model may change some of the analysis regarding the possible roles of HOS. For example, the most popular edge-based texture segmentation model is the Filter-Rectify-Filter (FRF) model, which consists in a V1-like filtering with rectification, followed by a second filtering stage capable of detecting texture-defined edges (Landy, 2013). Depending on the non-linearity used in such models (among other possible modifications), they may be able to find edges defined by HOS discontinuities, and these models have been shown to correlate with human HOS-based segmentation in central vision in one study (Zavitz & Baker, 2014). It is interesting to note that such a segmentation process could show sensitivity to HOS, even though still operating directly on rectified V1 outputs, instead of operating on units that encode HOS directly, such as V2 neuron outputs. This means that a simple segmentation model operating over V1 outputs could still explain some effect of HOS such as those observed in our experimental work (Herrera-Esposito et al., 2021). Another important edge-based segmentation mechanism is the emergence of selectivity to texture borders by tuned contextual modulation, which can emphasize the response of neurons near texture edges. This mechanism is proposed to be an important mechanism for computing segmentation and saliency in this area (Z. Li, 1999, 2002; Nothdurft et al., 2000). It is difficult to anticipate how HOS may affect these complex mechanisms when they operate on V1-like outputs, but they could lead to effects on segmentation that are not captured by our region-based segmentation task. Also, in another previous study (Ziemba et al., 2018), it is shown that contextual modulation for textures in V2 neurons is tuned to the HOS of the PS model, which could also give rise to this kind of segmentation based on contextual-modulation within V2. Nonetheless, see also (Schmid & Victor, 2014) where V2 neurons responses to texture-defined edges are argued to be compatible with a filter-rectify-filter mechanism, and less so with this kind of contextual modulation mechanism.

Acknowledgments

We thank Ruth Rosenholtz for useful comments on previous work that motivated this analysis. Ruben Coen-Cagli was supported in part by National Institutes of Health grant EY031166. Daniel Herrera-Esposito was supported by the studentships from the Comisión Académica de Posgrados, UdelaR, Uruguay, and by travelships awarded by PEDECIBA, UdelaR, Uruguay, and CSIC, Uruguay.

References

Adams, W. J., & Mamassian, P. (2004). Bayesian combination of ambiguous shape cues. *Journal of Vision*, *4*(10), 7–7. https://doi.org/10.1167/4.10.7

Bakin, J. S., Nakayama, K., & Gilbert, C. D. (2000). Visual Responses in Monkey Areas V1 and V2 to Three-Dimensional Surface Configurations. *Journal of Neuroscience*, *20*(21), 8188–8198. https://doi.org/10.1523/JNEUROSCI.20-21-08188.2000

Balas, B. (2006). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research*, *46*(3), 299–309. https://doi.org/10.1016/j.visres.2005.04.013

Balas, B. (2008). Attentive texture similarity as a categorization task: Comparing texture synthesis models. *Pattern Recognition*, *41*(3), 972–982. https://doi.org/10.1016/j.patcog.2007.08.007

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

Ben-Shahar, O. (2006). Visual saliency and texture segregation without feature gradient. *Proceedings of the National Academy of Sciences*, *103*(42), 15704–15709. https://doi.org/10.1073/pnas.0604410103

Bergen, J. R., & Landy, M. S. (1991). Computational Modeling of Visual Texture Segregation. In M. Landy & J. A. Movshon (Eds.), *Computational Models of Visual Processing* (pp. 253–271). MIT Press.

Bhatt, R., Carpenter, G. A., & Grossberg, S. (2007). Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research*, *47*(25), 3173–3211. https://doi.org/10.1016/j.visres.2007.07.013

Braddick, O. (1993). Segmentation versus integration in visual motion processing. *Trends in Neurosciences*, *16*(7), 263–268. https://doi.org/10.1016/0166-2236(93)90179-P

Brodatz, P. (1966). *Textures: A photographic album for artists and designers*. Dover Pubns.

Burge, J. (2020). Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science*, *6*, 22.1-22.27. https://doi.org/10.1146/annurev-vision-030320-041134

Cazettes, F., Fischer, B. J., & Pena, J. L. (2014). Spatial cue reliability drives frequency tuning in the barn Owl's midbrain. *ELife*, *3*, e04854. https://doi.org/10.7554/eLife.04854

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

Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision*, *16*(2), 13–13. https://doi.org/10.1167/16.2.13

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*(1), 146–154. https://doi.org/10.1038/nn.2983

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

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201. https://doi.org/10.1038/nn.2889

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981. https://doi.org/10.1038/nn.3402

Friedman, J., Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Noah Simon, & Junyang Qian. (2019). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models* (3.0-2) [Computer software]. https://rdrr.io/cran/glmnet/

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture Synthesis Using Convolutional Neural Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 262–270). Curran Associates, Inc. http://papers.nips.cc/paper/5633-texture-synthesis-using-convolutional-neural-networks.pdf

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724. https://doi.org/10.1016/S0042-6989(00)00277-7

Giora, E., & Casco, C. (2007). Region- and edge-based configurational effects in texture segmentation. *Vision Research*, *47*(7), 879–886. https://doi.org/10.1016/j.visres.2007.01.009

Gu, Y., Angelaki, D. E., & DeAngelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience*, *11*(10), 1201–1210. https://doi.org/10.1038/nn.2191

Hermundstad, A. M., Briguglio, J. J., Conte, M. M., Victor, J. D., Balasubramanian, V., & Tkačik, G. (2014). Variance predicts salience in central sensory processing. *ELife*, *3*, e03722. https://doi.org/10.7554/eLife.03722

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), 1–1. https://doi.org/10.1167/jov.21.1.1

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

Hindi Attar, C., Hamburger, K., Rosenholtz, R., Götzl, H., & Spillmann, L. (2007). Uniform versus random orientation in fading and filling-in. *Vision Research*, *47*(24), 3041–3051. https://doi.org/10.1016/j.visres.2007.07.022

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, *39*(21), 3621–3629. https://doi.org/10.1016/S0042-6989(99)00088-7

Julesz, B. (1962). Visual Pattern Discrimination. *IRE Transactions on Information Theory*, *8*(2), 84–92. https://doi.org/10.1109/TIT.1962.1057698

Julesz, B., Gilbert, E. N., & Victor, J. D. (1978). Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics*, *31*(3), 137–140.

https://doi.org/10.1007/BF00336998

Knierim, J. J., & van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology*, *67*(4), 961–980. https://doi.org/10.1152/jn.1992.67.4.961

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*(24), 2539–2558. https://doi.org/10.1016/S0042-6989(03)00458-9

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *Journal of Neuroscience*, *15*(2), 1605–1615. https://doi.org/10.1523/JNEUROSCI.15-02-01605.1995

Lamme, V. A. F., Rodriguez-Rodriguez, V., & Spekreijse, H. (1999). Separate Processing Dynamics for Texture Elements, Boundaries and Surfaces in Primary Visual Cortex of the Macaque Monkey. *Cerebral Cortex*, *9*(4), 406–413. https://doi.org/10.1093/cercor/9.4.406

Landy, M. S. (2013). Texture analysis and perception. In J.S. Werner & L.M. Chalupa (Eds.), *The new visual neurosciences* (pp. 639–652). MIT Press.

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research*, *31*(4), 679–691. https://doi.org/10.1016/0042-6989(91)90009-T

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1265–1278. https://doi.org/10.1109/TPAMI.2005.151

Li, A., & Lennie, P. (2001). Importance of color in the segmentation of variegated surfaces. *JOSA A*, *18*(6), 1240–1251. https://doi.org/10.1364/JOSAA.18.001240

Li, Z. (1999). Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Network: Computation in Neural Systems*, *10*(2), 187–212. https://doi.org/10.1088/0954-898X_10_2_305

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9–16. https://doi.org/10.1016/S1364-6613(00)01817-9

Machilsen, B., & Wagemans, J. (2011). Integration of contour and surface information in shape detection. *Vision Research*, *51*(1), 179–186. https://doi.org/10.1016/j.visres.2010.11.005

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *JOSA A*, *7*(5), 923–932. https://doi.org/10.1364/JOSAA.7.000923

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

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, *2*, 416–423 vol.2. https://doi.org/10.1109/ICCV.2001.937655

Møller, P., & Hurlbert, A. C. (1996). Psychophysical evidence for fast region-based segmentation processes in motion and color. *Proceedings of the National Academy of Sciences*, *93*(14), 7421–7426. https://doi.org/10.1073/pnas.93.14.7421

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic Depth: Its Relation to Image Segmentation, Grouping, and the Recognition of Occluded Objects. *Perception*, *18*(1), 55–68. https://doi.org/10.1068/p180055

Neri, P. (2014). Semantic Control of Feature Extraction from Natural Scenes. *Journal of Neuroscience*, *34*(6), 2374–2388. https://doi.org/10.1523/JNEUROSCI.1755-13.2014

Neri, P. (2017). Object segmentation controls image reconstruction from natural scenes. *PLOS Biology*, *15*(8), e1002611. https://doi.org/10.1371/journal.pbio.1002611

Nothdurft, H.-C., Gallant, J. L., & Van Essen, D. C. (2000). Response profiles to texture border patterns in area V1. *Visual Neuroscience*, *17*(3), 421–436. https://doi.org/10.1017/S0952523800173092

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences*, *112*(4), E351–E360. https://doi.org/10.1073/pnas.1415146112

Okazawa, G., Tajima, S., & Komatsu, H. (2017). Gradual Development of Visual Texture-Selective Properties Between Macaque Areas V2 and V4. *Cerebral Cortex*, *27*(10), 4867–4880. https://doi.org/10.1093/cercor/bhw282

Pavão, R., Sussman, E. S., Fischer, B. J., & Peña, J. L. (2020). Natural ITD statistics predict human auditory spatial perception. *ELife*, *9*, e51927. https://doi.org/10.7554/eLife.51927

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. https://doi.org/10.1023/A:1026553619983

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*.

Roelfsema, P. R. (2006). Cortical Algorithms for Perceptual Grouping. *Annual Review of Neuroscience*, *29*(1), 203–227. https://doi.org/10.1146/annurev.neuro.29.051605.112939

Rosenholtz, R. (2014). Texture perception. In *The Oxford Handbook of Perceptual Organization*. https://doi.org/10.1093/oxfordhb/9780199686858.013.058

Rosenholtz, R. (2016). Capabilities and Limitations of Peripheral Vision. *Annual Review of Vision Science*, *2*(1), 437–457. https://doi.org/10.1146/annurev-vision-082114-035733

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4), 14–14. https://doi.org/10.1167/12.4.14

Saarela, T. P., & Landy, M. S. (2012). Combination of texture and color cues in visual segmentation. *Vision Research*, *58*, 59–67. https://doi.org/10.1016/j.visres.2012.01.019

Salzburg texture image database (STex). (n.d.). http://wavelab.at/sources/STex/

Schmid, A. M., & Victor, J. D. (2014). Possible functions of contextual modulations and receptive field nonlinearities: Pop-out and texture segmentation. *Vision Research*, *104*, 57–67. https://doi.org/10.1016/j.visres.2014.07.002

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences*, *98*(4), 1935–1940. https://doi.org/10.1073/pnas.98.4.1935

Tesileanu, T., Conte, M. M., Briguglio, J. J., Hermundstad, A. M., Victor, J. D., & Balasubramanian, V. (2020). Efficient coding of natural scene statistics predicts discrimination thresholds for grayscale textures. *ELife*, *9*, e54347. https://doi.org/10.7554/eLife.54347

Tkacik, G., Prentice, J. S., Victor, J. D., & Balasubramanian, V. (2010). Local statistics in natural scenes predict the saliency of synthetic textures. *Proceedings of the National Academy of Sciences*, *107*(42), 18149–18154. https://doi.org/10.1073/pnas.0914916107

Vacher, J., & Coen-Cagli, R. (2019). Combining mixture models with linear mixing updates: Multilayer image segmentation and synthesis. *ArXiv:1905.10629 [Cs, q-Bio]*. http://arxiv.org/abs/1905.10629

Victor, J. D. (1994). Images, statistics, and textures: Implications of triple correlation uniqueness for texture statistics and the Julesz conjecture: comment. *JOSA A*, *11*(5), 1680–1684. https://doi.org/10.1364/JOSAA.11.001680

Victor, J. D., & Conte, M. M. (1996). The role of high-order phase correlations in texture processing. *Vision Research*, *36*(11), 1615–1631. https://doi.org/10.1016/0042-6989(95)00219-7

Victor, J. D., Conte, M. M., & Chubb, C. F. (2017). Textures as Probes of Visual Processing. *Annual Review of Vision Science*, *3*(1), 275–296. https://doi.org/10.1146/annurev-vision-102016-061316

Victor, J. D., Thengone, D. J., & Conte, M. M. (2013). Perception of second- and third-order orientation signals and their interactions. *Journal of Vision*, *13*(4), 21–21. https://doi.org/10.1167/13.4.21

Wagemans, J., Elder, J., Kubovy, M., Palmer, S., Peterson, M., Singh, M., & Heydt, R. von der. (2012). A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure–Ground Organization. *Psychological Bulletin*, *138*(6), 1172–1217. https://doi.org/10.1037/a0029333

Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, *17*(12), 5–5. https://doi.org/10.1167/17.12.5

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. https://doi.org/10.7554/eLife.42512

Wolfson, S. S., & Landy, M. S. (1998). Examining edge- and region-based texture analysis mechanisms. *Vision Research*, *38*(3), 439–446. https://doi.org/10.1016/S0042-6989(97)00153-3

Yu, Y., Schmid, A. M., & Victor, J. D. (2015). Visual processing of informative multipoint correlations arises primarily in V2. *ELife*, *4*, e06604. https://doi.org/10.7554/eLife.06604

Zavitz, E., & Baker, C. L. (2014). Higher order image structure enables boundary segmentation in the absence of luminance or contrast cues. *Journal of Vision*, *14*(4), 14–14. https://doi.org/10.1167/14.4.14

Zhang, X., Zhaoping, L., Zhou, T., & Fang, F. (2012). Neural Activities in V1 Create a Bottom-Up Saliency Map. *Neuron*, *73*(1), 183–192. https://doi.org/10.1016/j.neuron.2011.10.035

Zhaoping, L. (2019). A new framework for understanding vision from the perspective of the primary visual cortex. *Current Opinion in Neurobiology*, *58*, 1–10. https://doi.org/10.1016/j.conb.2019.06.001

Zhaoping, L., Guyader, N., & Lewis, A. (2009). Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection. *Journal of Vision*, *9*(11), 20–20. https://doi.org/10.1167/9.11.20

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149. https://doi.org/10.1073/pnas.1510847113

Ziemba, C. M., Freeman, J., Simoncelli, E. P., & Movshon, J. A. (2018). Contextual modulation of sensitivity to naturalistic image structure in macaque V2. *Journal of Neurophysiology*, *120*(2), 409–420. https://doi.org/10.1152/jn.00900.2017

Supplementary materials for: Redundancy between spectral and higher-order texture statistics for natural image segmentation

Daniel Herrera-Esposito[1]✱, Leonel Gómez-Sena[1], Ruben Coen-Cagli[2]

1 Laboratorio de Neurociencias, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
2 Dept. of Systems and Computational Biology and Dominick P. Purpura Dept. of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA
✱ Corresponding author. E-mail address: dherrera1911@gmail.com

Index:

S1) Non-linear and linear decoding show similar redundancy:

To test whether a non-linear decoder could extract further information from the statistics for segmentation, we analyzed whether the same patterns in performance were observed when neural networks were used to solve the natural image segmentation task.

We trained the neural networks using the Keras package for R (Francois Chollet, 2015) to classify the pairs of patches from the BSD, using the same training and testing division scheme as described for the linear classifier in the main text. We also normalized the differences in statistics of the pairs of the training set to have 0 mean and unit variance, and performed PCA on these values, keeping the principal components that explained 95% of the variance. This resulted in the number of principal components for each model shown in **Table S1**. Although we applied PCA simultaneously on all the statistics used by a given model, performance was similar when applying PCA on the subsets of statistics individually.

(SP1) To select the architecture and training regime of the network, we performed a search of parameters that maximized performance at the task. For this, we focused specifically on the performance of the model using pixel and HOS, which in preliminary analyses showed worse performance than the linear classifiers (suggesting overfitting of the nonlinear classifier). All networks had units with ReLu activation functions, and an output layer of 1 unit with a sigmoid activation function. Furthermore, all were trained using a binary cross entropy loss with the Adam gradient descent algorithm, with a batch size of 32. We tested 8 fully connected networks with the following architectures, with the hyphen separated numbers indicating the number of units in successive hidden layers in the network: 10, 30, 50, 10-2, 30-10, 50-20, 30-10-2, 50-20-5. Thus, we tested networks with 1 to 3 hidden layers, and different numbers of units. We also tested three different L1 regularization penalties: 0.001, 0.003 and 0.01. Furthermore, we trained the networks for a duration of 350 epochs. For each combination of parameters, 10 models were trained and tested on different random splits of the data into training and testing sets.

Statistics	**Principal components (95% variance)**
Pixel	8
Pixe + Spectral	45
Pixel + HOS	176
Pixel + Spectral + HOS	206
Spectral	38
HOS	169
Spectral + HOS	200

(SP2) Table S1. Number of principal components that retain 95% of the variance for each combination of statistics. The PCA was performed on all statistics in the indicated model together.

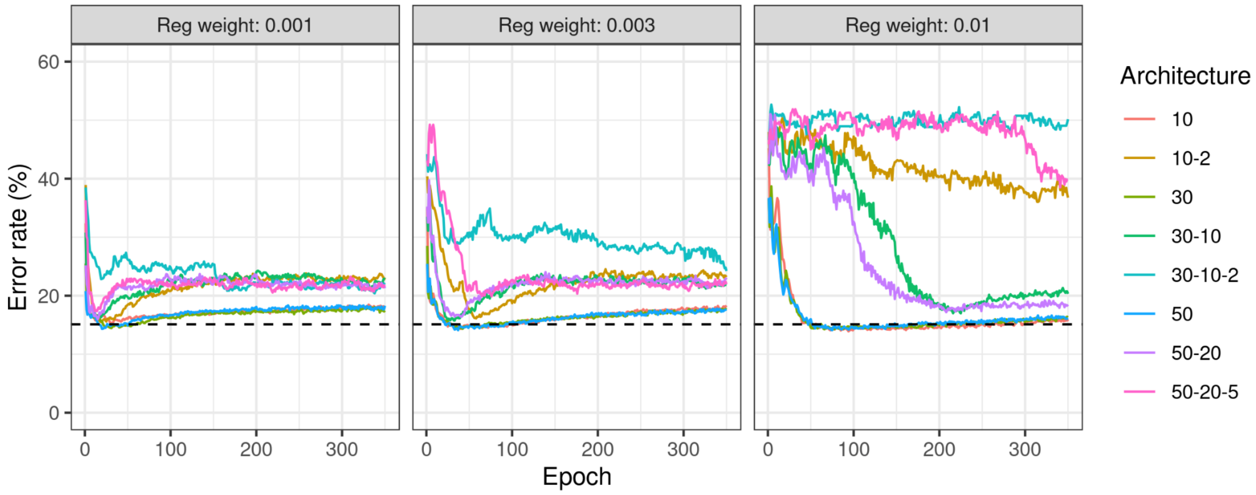


Figure S1. Segmentation performance during training for neural networks using HOS. Each line shows the average performance for 10 model instantiations using different train-validation set splits. The architecture of the model corresponding to each colored line is indicated in the legend to the right.

The horizontal dashed line shows the performance for the same set of statistics when using the linear model.

In **Figure S1** we see the performance throughout training for the different models using pixel and HOS. We observe that the networks that achieve the best performance are the simplest ones, consisting of only one hidden layer. Furthermore, these networks achieve best performance early in training, and then their performance deteriorates, probably due to overfitting. Also, we see that peak performance is similar to the performance of the linear model (indicated by the dashed horizontal line).

Therefore, following the results from the parameter exploration in **Figure S1**, we next tested a network with a single hidden layer of 30 units on all sets of statistics, using 80 epochs for training. The performance results shown in **Table S2** show that performance changes only slightly for all sets of statistics compared to the linear model.

Parameters	**Linear model:** **% error rate (SD)**	**Neural network:** **% error rate (SD)**
Pixel	20.7 (1.9)	20.0 (1.9)
Pixel + Spectral	16.8 (1.9)	16.3 (1.9)
Pixel + HOS	15.1 (1.8)	14.4 (1.7)
Pixel + Spectral + HOS	13.4 (1.7)	13.0 (1.7)
Spectral	17.6 (2.1)	17.0 (1.5)
HOS	19.5 (1.7)	18.4 (1.7)
Spectral + HOS	13.6 (1.5)	13.0 (1.3)

Table S2. Mean performance error in the segmentation task on natural image patches using each combination of texture statistics, with either the linear classifier shown in **Figure 3** in the main text, or using a neural network decoder with one hidden layer of 30 units. For each combination of statistics and each type of model, 20 instances of the model were trained with random splittings of the patches into train and test sets.

We note that although selecting the number of training epochs for which HOS performance was optimal may bias the resulting performances in favor of HOS, when analyzing the performance changes during training for spectral statistics and for the full model, the results were similar to those of HOS (data not shown).

S2) Spectral and HOS still perform similarly when matched in dimensionality:

Although the number of HOS in the PS model (552) is considerably larger than the number of spectral statistics in the PS model (137), both sets of statistics have considerable redundancy. Thus, we analyzed whether this difference in the number of statistics is maintained when performing PCA on these subsets. In **Table S3** we show the number of principal components (PC) required to retain different amounts of variance in the Berkeley Segmentation Dataset (BSD). We observe that the number of components needed to retain a fixed amount of variance was around 4 to 5 times larger for HOS than for spectral statistics in all cases. In **Table S4** we also see that the performance of the linear segmentation model using the PC for different levels of retained variance maintains roughly the same patterns as when using the original statistics space.

Statistics	Original number	95% variance	90% variance	80% variance	60% variance
Pixel	16	8	6	5	2
Spectral	137	38	24	12	4
HOS	552	169	112	61	23

Table S3. Number of principal components needed to retain the indicated percentages of variance of the pairs of BSD, for each combination of statistics.

Statistics	% error rate (all stats)	% error rate (95% variance PC)	% error rate (80% variance PC)	% error rate (60% variance PC)
Pixel	20.7	20.8	23.0	29.7
Pixel + Spectral	16.8	17.5	18.3	19.9
Pixel + HOS	15.1	16.3	17.2	20.6
Pixel + Spectral + HOS	13.4	13.8	14.7	16.5
Spectral	17.6	19.1	21.5	22.2
HOS	19.5	20.2	21.1	24.1
Spectral + HOS	13.6	14.5	16.4	17.9

Table S4. Segmentation performance in natural images using the PCs of each subset of statistics that retain the indicated amount of variance. Error rates show the average of 20 models trained on different random train-test splits.

Although this could be initially interpreted as spectral statistics offering similar performance with fewer parameters, when setting the number of PCs of HOS to be the same as for spectral statistics, the trends in performance across models do not change much, as shown in **Table S5.** This indicates that, at least roughly, spectral and HOS can perform segmentation to similar performances using similar numbers of dimensions.

Statistics	% error rate (all stats)	% error rate (95% variance PC)	% error rate (80% variance PC)	% error rate (60% variance PC)
Pixel	20.7	20.8	23.0	29.7
Pixel + Spectral	16.8	17.5	18.3	19.9
Pixel + HOS	15.1	17.4	20.0	23.1
Pixel + Spectral + HOS	13.4	13.9	15.4	16.5
Spectral	17.6	19.1	21.5	22.2
HOS	19.5	22.2	25.0	27.3
Spectral + HOS	13.6	15.2	17.5	18.5

Table S5. Segmentation performance in BSD using the components of spectral and pixel statistics that retain the indicated amount of variance, but fixing HOS to have the same number of PC as spectral statistics. Error rates show the average of 20 models trained on different random train-test splits.

S3) Models generalize poorly between scenes and texture datasets:

We hypothesized that due to the possible differences in the task of segmenting natural images (where cues other than texture are used to generate the labels) and texture segmentation, the results from scene segmentation may not be fully representative of the more specific process of texture segmentation in natural images. Since we found that our comparisons between the different kinds of statistics were qualitatively similar between the two, we wondered whether texture statistics may be useful in the same way for the two tasks. To answer this question, we tested whether the models trained in one dataset (i.e. BSD or our natural texture dataset) generalized to the other. For this, we used the same procedure for splitting each dataset into testing and training sets as described in the main text, but trained in the training set of one dataset, and tested in the other.

We observe in **Table S6** the performance results for these cross-trained models was considerably worse than for the models trained in the same dataset as they are tested (i.e. **Figure 3** and **Figure 4** in the main text), indicating that generalizability between the two datasets was poor.

Statistics	**% error rate in natural images (trained with textures)**	**% error rate natural Images (trained with natural images)**	**% error rate in textures (trained in natural images)**	**% error rate in textures (trained in textures)**
Pixel	35.6	20.7	37.9	24.5
Pixel + Spectral	28.8	16.8	21.4	12.7
Pixel + HOS	36.0	15.1	49.1	13.0
Pixel + Spectral + HOS	32.3	13.4	45.6	10.6
Spectral	27.6	17.6	20.2	15.4
HOS	33.4	19.5	50.6	14.5
Spectral + HOS	30.6	13.6	45.8	11.8

Table S6. Segmentation performance of models trained in one of the datasets and tested in the other. Error rates show the average of 20 models trained on different random train-test splits.

S4) Identification of pairs with useful HOS fails due to within-class heterogeneity:

We wondered whether we could interpret some of the reasons behind the low accuracy of the model trained to detect useful HOS. For this, we noted that there are some important characteristics of the data that is being fitted that could be related to this behavior. These characteristics are:

1) The “No HOS improvement” class is heterogeneous, containing pairs where: a) spectral and HOS are wrong, b) spectral statistics are correct, HOS are wrong, and c) spectral and HOS are both correct.

2) Each class in the dataset (“No HOS improvement”, “HOS improvement”) also has heterogeneity because they contain pairs where the segmentation ground truth is “matched” and where it is “unmatched”. That is, the improvement in classification by using HOS can be either because HOS show a small difference between the patches that favors no segmentation (when the ground truth is “matched”) or because HOS show a large difference that favors segmenting the patches (when the pair is “unmatched”).

3) The dataset is highly imbalanced, with only 10% of the pairs of patches being “improved by HOS”.

These properties of the dataset are due to the nature of the problem. 1) is because we want to find examples where HOS improves segmentation, 2) is because in order to have a classifier that tells us whether to use HOS or spectral statistics, and that is useful, we would want it to work without knowing the ground truth of the segmentation task, and 3) is due to how the classes are constructed, following 1). We hypothesize that it is these properties of the dataset (which follow from the nature of the problem) that make classification of HOS usefulness a difficult task.

Therefore, we propose that a control case in which our method for identifying pairs with useful HOS is expected to show high performance is when we apply the method to a subset of the data without characteristics 1), 2) and 3). To test this, we therefore trained the classifier on a subset of the data which has the following characteristics:

1) For the “Not improved” class we only use pairs of patches where spectral statistics are correct and HOS are incorrect. That is, we discard the pairs of patches where both statistics agree.

2) We only use pairs of patches where the ground truth segmentation is “unmatched”. This way, HOS usefulness is because it favors segmentation.

Because of the criteria described above, this subset of the data is also much more balanced across classes, due to the removal of datapoints from the “Not improved” class.

When performing the same procedure as in the manuscript for this subset of data we obtain an accuracy of 86%, much higher than for the complete dataset (**Table S7**).

	Label: No HOS improvement	**Label: HOS improvement**	
Predicted: No HOS improvement	157	28	185
Predicted: HOS improvement	16	118	134
	173	146	

Table S7. Classification of pairs of texture image patches as being better segmented by HOS or not, where the pairs of patches with agreement between spectral statistics and HOS and pairs with “matched” ground truth are not included in training and testing.

Interestingly, when we maintain the exclusion of pairs where both groups of statistics agree (criterion 1), but we remove the exclusion of "matched" pairs (criterion 2), performance drops steeply to 58% (**Table S8**). Thus, even the "simpler" task of deciding between spectral and HOS when they explicitly disagree is difficult to achieve with good accuracy. Note that this low accuracy is despite the classes in this example being more balanced.

	Label: No HOS improvement	**Label: HOS improvement**	
Predicted: No HOS improvement	218	135	353
Predicted: HOS improvement	118	135	253
	336	270	

Table S8. Classification of pairs of texture image patches as being better segmented by HOS or not, where the pairs of patches with agreement between spectral statistics and HOS not included in training and testing, but both "matched" and "unmatched" pairs are included.

Also, when we remove criterion 1 (we include pairs where spectral and HOS agree), but we keep criterion 2 (we remove the "matched" pairs), we also observe a decrease in performance, although less pronounced than in the previous case, with accuracy falling to 77% (**Table S9**). This indicates that the major source of the low accuracy is in the mix of "matched" and "unmatched" pairs, and not because of the heterogeneity of pairs with no HOS improvement, or due to the dataset imbalance.

	Label: No HOS improvement	**Label: HOS improvement**	
Predicted: No HOS improvement	875	38	913
Predicted: HOS improvement	259	108	367
	1134	146	

Table S9. Classification of pairs of texture image patches as being better segmented by HOS or not, where "matched" pairs are excluded from training and testing, but pairs of patches with agreement between the two sets of statistics are included.

Thus, with these examples we show a control case where the method of analysis of fitting a model to the model outputs is expected to succeed, and we find a high accuracy for this case. We also show that the poor performance of the model is mostly due to the presence in the dataset of examples where HOS improve the segmentation task by favoring segmentation of pairs (unmatched pairs), and examples where they improve segmentation by disfavoring segmentation of pairs (matched pairs). We also show that excluding the pairs where spectral and HOS agree improves classification in the cases where we only train and test on the pairs with “unmatched”, but that this manipulation alone is not sufficient to reach high performance.

S5) Low agreement between models of HOS fine-tuning and psychophysics:

One way to test the hypothesis that the visual system is fine-tuned to use HOS for segmentation in cases where they are more informative, is to identify or synthesize images where, from natural image statistics, we would predict HOS to be particularly useful for segmentation. Then, these could be used to probe human segmentation. Although such an empirical testing of this hypothesis is outside of the scope of this work, the results from our prior experimental work (Herrera-Esposito et al., 2021) allow for a preliminary exploratory analysis.

In this previous experimental work, despite observing an overall weak effect of HOS in segmentation, we observed some variation in the magnitude of the effect between the 4 different pairs of textures used. These were pairs of textures that had their spectral statistics matched, but that differed in their HOS. To test whether this variation between textures responds to a fine-tuning of the visual system to use HOS in some cases but not others, we analyzed whether the predictions from our segmentation models and from the models estimating usefulness of HOS for these pairs of textures relate to the observed variability in the psychophysical experiment.

For this, we first trained the linear model for segmentation on the BSD. Then, we tested the model on the pairs of textures used in our experimental work (Herrera-Esposito et al., 2021) that had matched spectral statistics but different HOS (experiment 3, Figure 6 in the original work), and extracted the linear activation from the model for each pair of textures. Because the models were trained with matched patches being the positive samples, we changed the sign of the linear output, so that higher (more positive) values indicate stronger segmentation. We repeated this procedure 20 times, because the 10-fold cross validation for choosing the scaling parameter (see above) adds some variability to the training procedure, and we report the mean and variation across models. The same procedure was also done using the models trained to identify pairs of patches with useful HOS. In this case, we did not have to change the sign of the linear output, since the positive samples were those with useful HOS. The same procedure for both kinds of models was also reproduced using the natural texture dataset instead of the BSD.

For the experimental data, we extracted the effect fitted to each texture using generalized linear mixed models, as described in the experimental study. Specifically, we extracted the interaction term between HOS dissimilarity and texture discontinuity from the full model, fitted to each texture individually. This fitted term is in units of log-odds ratio, and because of the coding of variables used in the statistical model, we also changed the sign so that larger (more positive) values indicated stronger segmentation (see original work for more detail).

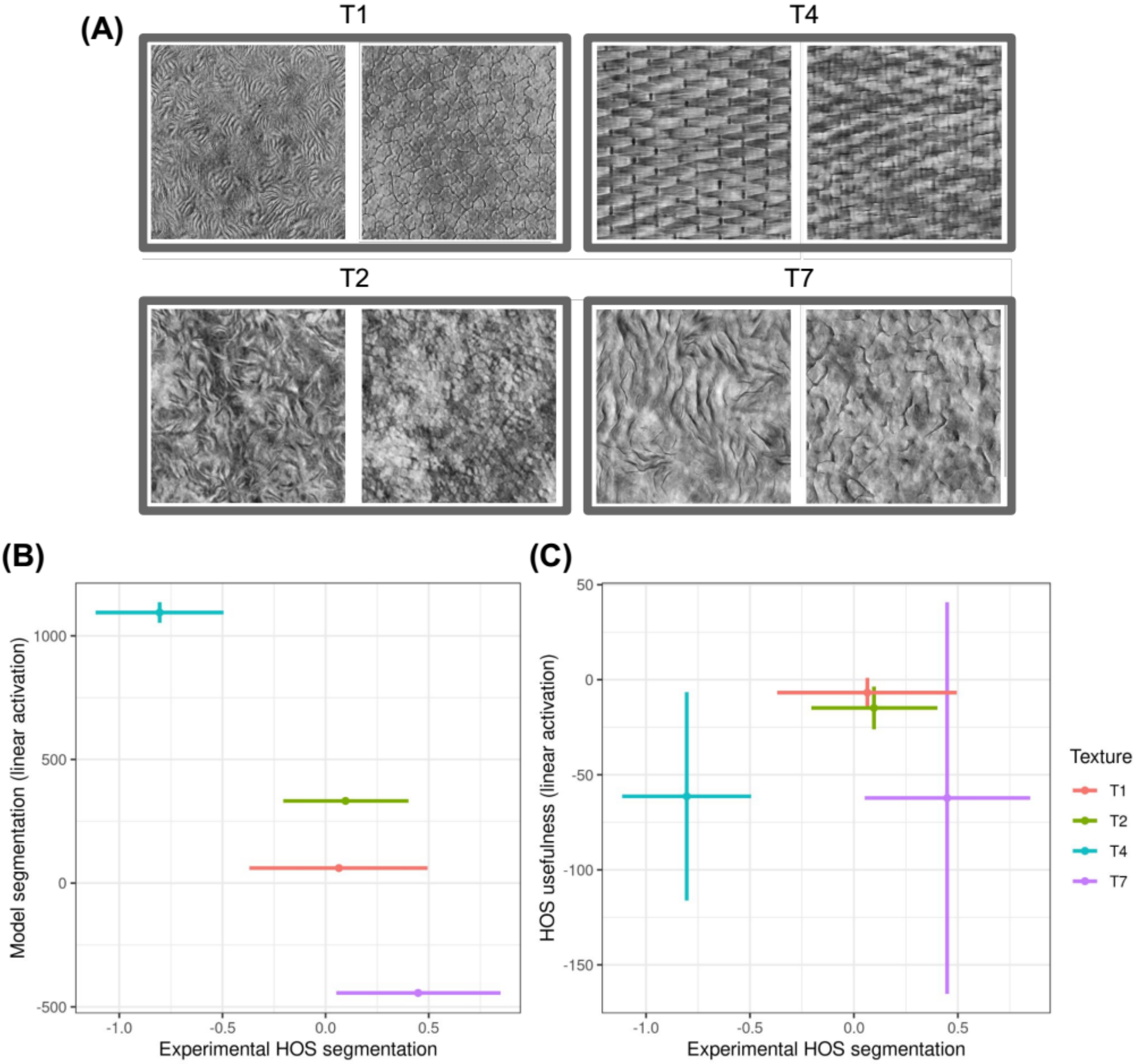


Figure S2. Comparison of experimental texture segmentation results (data obtained from (Herrera-Esposito et al., 2021)). The horizontal axis shows the estimated effect, on the segmentation of two patches of texture, of introducing a HOS mismatch between the two. Larger values indicate stronger segmentation when inducing the HOS mismatch. **(A)** Textures used in the experiment. Each

gray rectangle groups together two patches of the textures that comprise a pair. **(B)** The vertical axis shows the linear activation component of the segmentation model using only HOS, trained on natural images and then tested on the textures of the experimental stimuli (see Methods). **(C)** The vertical axis shows the linear activation of the model trained to identify pairs of natural image patches where HOS are better than spectral statistics for segmentation. The texture numbering of the cited work is maintained in this figure, and all 4 textures used are shown.

In **Figure S2A** we show, for each pair of textures, the estimated empirical effect of HOS dissimilarity in human segmentation, and the linear component of the segmentation model using only the image HOS. We see that the strength of the segmentation shown by the model does not follow the observed experimental segmentation. This is the case also if we remove texture T4, which we argue to be an outlier in the original work, where we attribute the strong negative effect of HOS dissimilarity on segmentation to an artifact due to a phase effect (see further discussion in (Herrera-Esposito et al., 2021), Experiment 3). In **Figure S2B** we show the same analysis, using the linear component for the model predicting whether HOS perform better than spectral statistics (i.e. whether HOS should be used to segment the image). Again, we see that there is hardly any clear relation between experimental segmentation and model prediction. These results are similar when using models trained on textures, rather than natural images, and also when using both spectral and HOS together as predictors in the models.

Due to the small number of textures used in this analysis, the uncertainty in the estimates of the individual textures, and the fact that the experiment was not designed to test this hypothesis, this result should be taken with great caution. But the lack of a clear association between the observed human segmentation and the segmentation from models using HOS, or the estimate of HOS usefulness, may reflect that these HOS are a weak segmentation cue overall in peripheral vision.

6) CLOSING REMARKS

In Chapter 1, I introduced the main topics related to this thesis, together with a brief introduction to the approach used in this work. In Chapter 2, I gave a broad overview and outlined the history of different sub-fields of vision science which have had many interactions in the recent neuroscience literature, and at whose intersection this thesis belongs. In Chapter 3, I summarized the main contributions of this thesis in the context of these sub-fields and their recent interactions, as well as future work. Chapters 4 and 5 present the results of the thesis and in-depth discussion in the context of the latest developments in the related literature. In this closing chapter, I provide a more informal discussion of how this work compares to the original goals and hypotheses (to the best of my certainly imperfect recollection capabilities), and the overarching questions that I find of particular interest at the end of this project, which are not necessarily the same as when I started.

This PhD project started with the goal of studying the phenomenon of texture filling-in in peripheral vision. In the first year, I learned about PS textures, about crowding, and about the summary-statistics model of peripheral vision. Then, due to experimental difficulties in studying filling-in, I changed the question to whether textures suffered from crowding, given that there were no previous reports of this, and there were several links between crowding and texture perception. During this process, I phenomenally observed that texture segmentation was a major determinant of contextual modulation, learned about the uncrowding literature, and decided that PS textures were an ideal kind of stimuli to study that phenomenon. Then, I also learned about the role of natural image statistics in contextual modulation, and decided to also try out connecting this research plan to natural image statistics. Finally, I learned about the so-called reproducibility crisis in psychology, and how difficult, yet important, it was to collect large-enough samples for my experiments, and to analyze the results properly.

After a lot of data collecting and statistical analysis, many of the questions I had originally asked were answered. Some were in line with what I expected:

segmentation cues reduced contextual modulation, and that was dependent on target-flanker continuity. We used this result to argue for the importance of recurrence and segmentation in early visual processing in the peripheral vision. Other results were more unexpected: differences in HOS did not induce strong segmentation in peripheral vision, and contextual modulation of textures in the periphery is not clearly dominated by crowding in our experiments. Also, I collected as much data as I could manage with the available resources, and even if costly, and I always chose to stretch the work to gather more data if I considered it necessary to have solid results. One example of this was the implementation of the SS observer model in the first paper, which was a response to Corey Ziemba's thoughtful challenges to my interpretations of the experimental results, which were originally based on intuition and word models. Product of this, I think that the results and analyses in this work are very robust, even in my interpretations of them may be challenged. Also product of this, the first stage of this work took longer than I would have wished, given my desire to tackle new questions.

After this first experimental stage, I had many interesting answers, and new questions, but also had one problem. Given that the early work was carried out without one single, clear question in mind, it was very difficult to put all that had been done together into one story. It was possible to center the work around two major points: that segmentation effects in our model could not be explained by feedforward SS representations, or that crowding may not be the limiting factor in texture perception in the periphery, with implications for the importance of crowding for natural vision, but it was difficult to put it all into one story. After much writing and editing work[1], and with Ruben's immense help, we managed to get a straight story out of the experiments. The product is the Journal of Vision paper of Chapter 4. We managed to respond to several of the questions I originally found of interest, and we connected them to vast pieces of the vision sciences literature. I hope that these results and discussions can serve as a solid experimental ground in which to base future studies into the physiology of the primate visual system, the improvement of both classical models of the early visual system and more recent neural network models, and future perceptual studies on the functions and roles of peripheral vision.

[1] A funny anecdote of this process that illustrates the messiness of writing this paper is that at one stage, with the manuscript almost completely written, all figures that are in the supplementary in the published version were actually in the main text, and vice versa, and they were all interchanged in one swift edit.

The results from the first study raise many questions, of which I will mention two I find particularly interesting. One important question pertains to the physiological explanations of why the HOS of the PS model induce only weak segmentation. The representation of these HOS mostly arises in mid-level visual areas V2 and V4, while that of spectral statistics known to induce segmentation arise in V1. One simple explanation is that feature or statistics based segmentation occurs only on visual information explicitly represented in V1. This would have important implications for the modeling of the visual system, particularly in the current age of neural networks that use a series of modules where the same computations are repeated. Are there important differences between the computations involved in going from early to mid-level visual areas and those involved in going from mid-level to higher level areas that make the former but not the latter support segmentation? As said above, this would have important implications for computational models of vision. On the other hand, may these experimental results be the product of the stimuli used, where the spectral statistics of the textures were matched, maybe leading to only small differences in the activity patterns generated by the two textures that hinder segmentation? If so, what implications does this have for the specific representations of summary statistics in mid-level visual areas in the periphery? For this last question, it is of particular interest to consider that our textures differing in HOS and matched in spectral statistics have different appearances under foveal vision, and that HOS dissimilarity had a considerable effect on our model observer. Are these SS representations dominated by relatively low-level descriptions that do not capture well the difference between the textures used in these experiments? Also, we observed that phase-scrambling the surrounds strongly reduced contextual modulation, and that this is likely mediated by segmentation. How can HOS dissimilarity not induce segmentation, but at the same time the lack of HOS be so important? Although there is no clear picture of what neural computations in the early visual system may give rise to these patterns, these results using stimuli and a model tightly linked to physiology offer rich grounds for generating hypothesis about processing in the early to mid-level visual system that can be tested with techniques for measuring neural activity.

Also, interestingly, we find that crowding may not be the main limitation of texture perception in peripheral vision. This leads to the discussion at the end of Chapter 4. The more general question to which our observation relates is: what visual information is important for peripheral vision function? There is a vast literature dedicated to studying identification of objects, letters, and artificial stimuli in the peripheral vision, and what aspects of peripheral vision limit these. This has led crowding, which interferes with target identification, to be labeled the most important limitation of peripheral vision. But, what if identification of stimuli plays only a minor role in the everyday workings of peripheral vision? What if most of the tasks we perform with peripheral vision, such as guiding navigation and movement, spatial perception, saliency detection, scene segmentation, and others, have little to do with object recognition? Would the striking relevance of visual crowding then be just the product of studying peripheral vision using the wrong tasks? These ideas, although not further developed experimentally in this thesis, relate to new important directions that the study of the brain is undertaking. It is increasingly argued that neuroscience and cognitive science have put too little emphasis in past decades in the analysis of complex behaviors, such as those where action and perception interact. Although the roads taken by neuroscience and cognitive science in those times has been very fruitful and have led to striking advances, it is important to wonder what blind spots (no pun intended) may exist in this body of research. I think that an interesting example of this is crowding research, where a big emphasis has been placed in object recognition in the periphery that may not be representative of the actual limitations of peripheral vision. Whether our finding that texture contextual modulation does not show clear signature of crowding is a general phenomenon related to texture processing in the periphery, or the product of our specific task design, as suggested by Corey Ziemba in previous discussions (in which he was at least partly right) remains to be determined. But maybe it is not a coincidence that texture perception was a lengthily discussed subject by Gibson, the father of ecological perception and of the proposal of studying perception in its interaction with action, behavior and the environment, and we should better study what behaviors can be supported by texture information, and whether textures escape peripheral vision limitations to a larger degree than objects.

Finally, as the experimental work was finished, I pursued one of the venues I had grown an interest in, the analysis of natural image statistics. However, I did not study natural statistics in the problem that originally got me interested in this topic, contextual modulation, but on the problem of image segmentation. This problem arised from a common sense explanation of the small role we observed experimentally for HOS dissimilarities in texture segmentation. This analysis thus had the strong appeal of testing an hypothesis about the natural world from experimental observations of human perception. More interestingly, few studies to my knowledge compare models performing this relatively high level visual tasks with natural images to human perception, which is also a venue of research suggested to be in need of further development. As described in Chapter 5, we showed using an analysis of natural images that the HOS of the PS model add little to spectral statistics for the task of segmentation. We also propose that the apparent discrepancy between the importance of HOS of the PS model for texture perception and for texture segmentation may be explained by a varying usefulness of these statistics across tasks. According to this hypothesis, that we expect these statistics to add more information to texture identification, or material perception tasks. Preliminary analysis in these directions have shown promising results. A very ambitious extension of this line of work could relate to the questions left open in the previous paragraph. An extension to tasks such as identification of spatial layout in scenes or controlling a moving agent, and an extension of the information to a description of local texture across the scene, with a fovea-periphery layout could lead to a better understanding of why we see like we see in peripheral vision.

7) REFERENCES

Adelson, Edward H. 2001. “On Seeing Stuff: The Perception of Materials by Humans and Machines.” *Proceedings of the SPIE* 4299: 1–12. https://doi.org/10.1117/12.429489.

Adelson, Edward H., and James R. Bergen. 1985. “Spatiotemporal Energy Models for the Perception of Motion.” *JOSA A* 2 (2): 284–99. https://doi.org/10.1364/JOSAA.2.000284.

Adelson, Edward H, and James R Bergen. 1991. “The Plenoptic Function and the Elements of Early Vision.” In *Computational Models of Visual Processing*, 2:3–20. MIT Press.

Angelucci, Alessandra, Maryam Bijanzadeh, Lauri Nurminen, Frederick Federer, Sam Merlin, and Paul C. Bressloff. 2017. “Circuits and Mechanisms for Surround Modulation in Visual Cortex.” *Annual Review of Neuroscience* 40 (1): 425–51. https://doi.org/10.1146/annurev-neuro-072116-031418.

Angelucci, Alessandra, Jonathan B Levitt, Emma J S Walton, Jean-Michel Hupe, Jean Bullier, and Jennifer S Lund. 2002. “Circuits for Local and Global Signal Integration in Primary Visual Cortex.” *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 22 (19): 8633–46.

Anzai, Akiyuki, Xinmiao Peng, and David C Van Essen. 2007. “Neurons in Monkey Visual Area V2 Encode Combinations of Orientations.” *Nature Neuroscience* 10 (10): 1313–21. https://doi.org/10.1038/nn1975.

Balas, Benjamin, Lisa Nakano, and Ruth Rosenholtz. 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.

Barlow, Horace. 2001. “Redundancy Reduction Revisited.” *NETWORK: COMPUTATION IN NEURAL SYSTEMS* 12: 241–53.

Barth, Erhardt, Christoph Zetzsche, and Ingo Rentschler. 1998. “Intrinsic Two-Dimensional Features as Textons.” *JOSA A* 15 (7): 1723–32. https://doi.org/10.1364/JOSAA.15.001723.

Beck, Jacob. 1966. “Effect of Orientation and of Shape Similarity on Perceptual Grouping.” *Perception & Psychophysics* 1 (5): 300–302. https://doi.org/10.3758/BF03207395.

Beck, Jacob, K. Prazdny, and Azriel Rosenfeld. 1983. “A Theory of Textural Segmentation.” In *Human and Machine Vision*, edited by Jacob Beck, Barbara Hope, and Azriel Rosenfeld, 1–38. Notes and Reports in Computer Science and Applied Mathematics. Academic Press. https://doi.org/10.1016/B978-0-12-084320-6.50007-4.

Beck, Jacob, Anne Sutter, and Richard Ivry. 1987. “Spatial Frequency Channels and Perceptual Grouping in Texture Segregation.” *Computer Vision, Graphics, and Image Processing* 37 (2): 299–325. https://doi.org/10.1016/S0734-189X(87)80006-3.

Bell, Anthony J., and Terrence J. Sejnowski. 1997. “The ‘independent Components’ of Natural Scenes Are Edge Filters.” *Vision Research* 37 (23): 3327–38. https://doi.org/10.1016/S0042-6989(97)00121-1.

Bergen, James R., and Edward H. Adelson. 1988. “Early Vision and Texture Perception.” *Nature* 333 (6171): 363–64. https://doi.org/10.1038/333363a0.

Bergen, James R, and Michael S Landy. 1991. “Computational Modeling of Visual Texture Segregation.” In *Computational Models of Visual Processing*, edited by M. Landy and J. A. Movshon, 253–71. MIT Press.

Bouma, H. 1970. “Interaction Effects in Parafoveal Letter Recognition.” *Nature* 226 (5241): 177–78. https://doi.org/10.1038/226177a0.

Burge, Johannes. 2020. “Image-Computable Ideal Observers for Tasks with Natural Stimuli.” *Annual Review of Vision Science* 6 (September): 22.1-22.27. https://doi.org/10.1146/annurev-vision-030320-041134.

Carandini, Matteo, and David J. Heeger. 2012. “Normalization as a Canonical Neural Computation.” *Nature Reviews Neuroscience* 13 (1): 51–62. https://doi.org/10.1038/nrn3136.

Coen-Cagli, Ruben, Peter Dayan, and Odelia Schwartz. 2012. “Cortical Surround Interactions and Perceptual Salience via Natural Scene Statistics.” *PLOS Computational Biology* 8 (3): e1002405. https://doi.org/10.1371/journal.pcbi.1002405.

Coen-Cagli, Ruben, Adam Kohn, and Odelia Schwartz. 2015. “Flexible Gating of Contextual Influences in Natural Vision.” *Nature Neuroscience* 18 (11): 1648–55. https://doi.org/10.1038/nn.4128.

Cohen, Michael A., Daniel C. Dennett, and Nancy Kanwisher. 2016. “What Is the Bandwidth of Perceptual Experience?” *Trends in Cognitive Sciences* 20 (5): 324–35. https://doi.org/10.1016/j.tics.2016.03.006.

Curcio, Christine A., and Kimberly A. Allen. 1990. “Topography of Ganglion Cells in Human Retina.” *Journal of Comparative Neurology* 300 (1): 5–25. https://doi.org/10.1002/cne.903000103.

De Valois, R L, and K K De Valois. 1980. “Spatial Vision.” *Annual Review of Psychology* 31 (1): 309–41. https://doi.org/10.1146/annurev.ps.31.020180.001521.

Desimone, Robert, Thomas D. Albright, Charles G. Gross, and Charles Bruce. 1984. “Stimulus-Selective Properties of Inferior Temporal Neurons in the Macaque.” *The Journal of Neuroscience* 4 (8): 12.

DiCarlo, James J., Davide Zoccolan, and Nicole C. Rust. 2012. “How Does the Brain Solve Visual Object Recognition?” *Neuron* 73 (3): 415–34. https://doi.org/10.1016/j.neuron.2012.01.010.

Doerig, Adrien, A. Bornet, O. H. Choung, and M. H. Herzog. 2020. “Crowding Reveals Fundamental Differences in Local vs. Global Processing in Humans and Machines.” *Vision Research* 167 (February): 39–45. https://doi.org/10.1016/j.visres.2019.12.006.

Doerig, Adrien, Alban Bornet, Ruth Rosenholtz, Gregory Francis, Aaron M. Clarke, and Michael H. Herzog. 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, Adrien, Lynn Schmittwilken, Bilge Sayim, Mauro Manassi, and Michael H. Herzog. 2020. “Capsule Networks as Recurrent Models of Grouping and Segmentation.” *PLOS Computational Biology* 16 (7): e1008017. https://doi.org/10.1371/journal.pcbi.1008017.

Ehinger, Krista A., and Ruth Rosenholtz. 2016. “A General Account of Peripheral Encoding Also Predicts Scene Perception Performance.” *Journal of Vision* 16 (2): 13–13. https://doi.org/10.1167/16.2.13.

Farzin, Faraz, Susan M. Rivera, and David Whitney. 2009. “Holistic Crowding of Mooney Faces.” *Journal of Vision* 9 (6): 18–18. https://doi.org/10.1167/9.6.18.

Felleman DJ and Van Essen DC. 1991. “Distributed Hierarchical Processing in the Primate Cerebral Cortex.” *Cerebral Cortex (New York, N.Y. : 1991)* 1 (1): 1–47. https://doi.org/10.1093/cercor/1.1.1.

Field, David J. 1987. “Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells.” *Journal of the Optical Society of America A* 4 (12): 2379. https://doi.org/10.1364/JOSAA.4.002379.

Francis, Gregory, Mauro Manassi, and Michael H. Herzog. 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, Jeremy, and Eero P. Simoncelli. 2011. “Metamers of the Ventral Stream.” *Nature Neuroscience* 14 (9): 1195–1201. https://doi.org/10.1038/nn.2889.

Freeman, Jeremy, Corey M. Ziemba, David J. Heeger, Eero P. Simoncelli, and J. Anthony Movshon. 2013. “A Functional and Perceptual Signature of the Second Visual Area in Primates.” *Nature Neuroscience* 16 (7): 974–81. https://doi.org/10.1038/nn.3402.

Frisby, J., and James V. Stone. 2010. *Seeing : The Computational Approach to Biological Vision*. Vol. 88. Cambridge, Mass: MIT Press.

Gattass, Ricardo, Sheila Nascimento-Silva, Juliana G.M Soares, Bruss Lima, Ana Karla Jansen, Antonia Cinira M Diogo, Mariana F Farias, et al. 2005. “Cortical Visual Areas in Monkeys: Location, Topography, Connections, Columns, Plasticity and Cortical Dynamics.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 360 (1456): 709–31. https://doi.org/10.1098/rstb.2005.1629.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. 2016. “Image Style Transfer Using Convolutional Neural Networks.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2414–23. Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CVPR.2016.265.

Gatys, Leon, Alexander S Ecker, and Matthias Bethge. 2015. “Texture Synthesis Using Convolutional Neural Networks.” In *Advances in Neural Information Processing Systems 28,* edited by C. Cortes, N. D. Lawrence, D. D. Lee, M.

Sugiyama, and R. Garnett, 262–70. Curran Associates, Inc. http://papers.nips.cc/paper/5633-texture-synthesis-using-convolutional-neural-networks.pdf.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2018. “ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness.” *ArXiv:1811.12231 [Cs, q-Bio, Stat]*, November. http://arxiv.org/abs/1811.12231.

Geisler, Wilson S. 2008. “Visual Perception and the Statistical Properties of Natural Scenes.” *Annual Review of Psychology* 59 (1): 167–92. https://doi.org/10.1146/annurev.psych.58.110405.085632.

Gibson, James J. 1958. “Visually Controlled Locomotion and Visual Orientation in Animals.” *British Journal of Psychology* 49 (3): 182–94. https://doi.org/10.1111/j.2044-8295.1958.tb00656.x.

Gilbert, C. D., A. Das, M. Ito, M. Kapadia, and G. Westheimer. 1996. “Spatial Integration and Cortical Dynamics.” *Proceedings of the National Academy of Sciences* 93 (2): 615–22. https://doi.org/10.1073/pnas.93.2.615.

Gong, Mingliang, Yuming Xuan, L. James Smart, and Lynn A. Olzak. 2018. “The Extraction of Natural Scene Gist in Visual Crowding.” *Scientific Reports* 8 (1): 1–13. https://doi.org/10.1038/s41598-018-32455-6.

Goodale, Melvyn A., and A. David Milner. 1992. “Separate Visual Pathways for Perception and Action.” *Trends in Neurosciences* 15 (1): 20–25. https://doi.org/10.1016/0166-2236(92)90344-8.

Graham, Norma, Anne Sutter, and Charu Venkatesan. 1993. “Spatial-Frequency- and Orientation-Selectivity of Simple and Complex Channels in Region Segregation.” *Vision Research* 33 (14): 1893–1911. https://doi.org/10.1016/0042-6989(93)90017-Q.

Guo, Kun, Robert G. Robertson, Sasan Mahmoodi, and Malcolm P. Young. 2005. “Centre-Surround Interactions in Response to Natural Scene Stimulation in the Primary Visual Cortex.” *European Journal of Neuroscience* 21 (2): 536–48. https://doi.org/10.1111/j.1460-9568.2005.03858.x.

Heeger, David J, and James R Bergen. 1995. “Pyramid-Based Texture Analysis/Synthesis.” In *Proceedings of the 22nd Annual Conference on*

Computer Graphics and Interactive Techniques, 229–38. SIGGRAPH ’95. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/218380.218446.

Hegdé, Jay, and David C. Van Essen. 2000. “Selectivity for Complex Shapes in Primate Visual Area V2.” *Journal of Neuroscience* 20 (5): RC61–RC61. https://doi.org/10.1523/JNEUROSCI.20-05-j0001.2000.

Henry, C. A., S. Joshi, D. Xing, R. M. Shapley, and M. J. Hawken. 2013. “Functional Characterization of the Extraclassical Receptive Field in Macaque V1: Contrast, Orientation, and Temporal Dynamics.” *Journal of Neuroscience* 33 (14): 6230–42. https://doi.org/10.1523/JNEUROSCI.4155-12.2013.

Henry, Christopher A, Mehrdad Jazayeri, Robert M Shapley, and Michael J Hawken. 2020. “Distinct Spatiotemporal Mechanisms Underlie Extra-Classical Receptive Field Modulation in Macaque V1 Microcircuits.” Edited by Kristine Krug, Andrew J King, and Kristine Krug. *ELife* 9 (May): e54264. https://doi.org/10.7554/eLife.54264.

Hermundstad, Ann M, John J Briguglio, Mary M Conte, Jonathan D Victor, Vijay Balasubramanian, and Gašper Tkačik. 2014. “Variance Predicts Salience in Central Sensory Processing.” Edited by Timothy Behrens. *ELife* 3 (November): e03722. https://doi.org/10.7554/eLife.03722.

Herrera-Esposito, Daniel, Ruben Coen-Cagli, and Leonel Gomez-Sena. 2021. “Flexible Contextual Modulation of Naturalistic Texture Perception in Peripheral Vision.” *Journal of Vision* 21 (1): 1–1. https://doi.org/10.1167/jov.21.1.1.

Herrera-Esposito, Daniel, Leonel Gómez-Sena, and Ruben Coen-Cagli. 2021. “Redundancy between Spectral and Higher-Order Texture Statistics for Natural Image Segmentation.” *Vision Research* 187 (October): 55–65. https://doi.org/10.1016/j.visres.2021.06.007.

Herzog, Michael H., and Mauro Manassi. 2015. “Uncorking the Bottleneck of Crowding: A Fresh Look at Object Recognition.” *Current Opinion in Behavioral Sciences*, Cognitive control, 1 (February): 86–93. https://doi.org/10.1016/j.cobeha.2014.10.006.

Herzog, Michael H., Bilge Sayim, Vitaly Chicherov, and Mauro Manassi. 2015. "Crowding, Grouping, and Object Recognition: A Matter of Appearance." *Journal of Vision* 15 (6): 5–5. https://doi.org/10.1167/15.6.5.

Hubel, D. H., and T. N. Wiesel. 1959. "Receptive Fields of Single Neurones in the Cat's Striate Cortex." *The Journal of Physiology* 148 (3): 574–91. https://doi.org/10.1113/jphysiol.1959.sp006308.

Hubel, D. H., and T. N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160 (1): 106–54. https://doi.org/10.1113/jphysiol.1962.sp006837.

Hubel, D. H., and T. N. Wiesel. 1968. "Receptive Fields and Functional Architecture of Monkey Striate Cortex." *The Journal of Physiology* 195 (1): 215–43. https://doi.org/10.1113/jphysiol.1968.sp008455.

Hyvärinen, Aapo, Jarmo Hurri, and Patrick O. Hoyer. 2009. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* Springer Science & Business Media.

Ikeda, Hanako, Katsumi Watanabe, and Patrick Cavanagh. 2013. "Crowding of Biological Motion Stimuli." *Journal of Vision* 13 (4): 20–20. https://doi.org/10.1167/13.4.20.

Julesz, B. 1962. "Visual Pattern Discrimination." *IRE Transactions on Information Theory* 8 (2): 84–92. https://doi.org/10.1109/TIT.1962.1057698.

Julesz, B., E. N. Gilbert, L. A. Shepp, and H. L. Frisch. 1973. "Inability of Humans to Discriminate between Visual Textures That Agree in Second Order Statistics: Revisited." *Perception* 2 (4): 391–405. https://doi.org/10.1068/p020391.

Julesz, B., E. N. Gilbert, and J. D. Victor. 1978. "Visual Discrimination of Textures with Identical Third-Order Statistics." *Biological Cybernetics* 31 (3): 137–40. https://doi.org/10.1007/BF00336998.

Julesz, Bela. 1962. "Visual Pattern Discrimination." *IRE Transactions on Information Theory* 8 (2): 84–92. https://doi.org/10.1109/TIT.1962.1057698.

Krieger, G., and C. Zetzsche. 1996. "Nonlinear Image Operators for the Evaluation of Local Intrinsic Dimensionality." *IEEE Transactions on Image Processing* 5 (6): 1026–42. https://doi.org/10.1109/83.503917.

LaLonde, Rodney, and Ulas Bagci. 2018. "Capsules for Object Segmentation." *ArXiv:1804.04241 [Cs, Stat]*, April. http://arxiv.org/abs/1804.04241.

Landy, Michael S. 2013. “Texture Analysis and Perception.” In *The New Visual Neurosciences*, edited by J.S. Werner and L.M. Chalupa, 639–52. MIT Press.

Lettvin, J. Y., H. R. Maturana, H. R. Maturana, W. S. Mcculloch, and W. H. Pitts. 1959. “What the Frog’s Eye Tells the Frog’s Brain.” *Proceedings of the IRE* 47 (11): 1940–51. https://doi.org/10.1109/JRPROC.1959.287207.

Lettvin, Jerome Y. 1976. “On Seeing Sidelong.” *The Sciences* 16 (4): 10–20. https://doi.org/10.1002/j.2326-1951.1976.tb01231.x.

Levi, Dennis M. 2008. “Crowding—An Essential Bottleneck for Object Recognition: A Mini-Review.” *Vision Research* 48 (5): 635–54. https://doi.org/10.1016/j.visres.2007.12.009.

Levi, Dennis M., Stanley A. Klein, and A.P. Aitsebaomo. 1985. “Vernier Acuity, Crowding and Cortical Magnification.” *Vision Research* 25 (7): 963–77. https://doi.org/10.1016/0042-6989(85)90207-X.

Louie, Elizabeth G., David W. Bressler, and David Whitney. 2007. “Holistic Crowding: Selective Interference between Configural Representations of Faces in Crowded Scenes.” *Journal of Vision* 7 (2): 24–24. https://doi.org/10.1167/7.2.24.

Malania, Maka, Michael H. Herzog, and Gerald Westheimer. 2007. “Grouping of Contextual Elements That Affect Vernier Thresholds.” *Journal of Vision* 7 (2): 1–1. https://doi.org/10.1167/7.2.1.

Malik, Jitendra, and Pietro Perona. 1990. “Preattentive Texture Discrimination with Early Vision Mechanisms.” *JOSA A* 7 (5): 923–32. https://doi.org/10.1364/JOSAA.7.000923.

Manassi, Mauro, Frouke Hermens, Gregory Francis, and Michael H. Herzog. 2015. “Release of Crowding by Pattern Completion.” *Journal of Vision* 15 (8): 16–16. https://doi.org/10.1167/15.8.16.

Manassi, Mauro, Sophie Lonchampt, Aaron Clarke, and Michael H. Herzog. 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, Mauro, Bilge Sayim, and Michael H. Herzog. 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, Mauro, Bilge Sayim, and Michael H. Herzog. 2013. “When Crowding of Crowding Leads to Uncrowding.” *Journal of Vision* 13 (13): 10–10. https://doi.org/10.1167/13.13.10.

Marĉelja, S. 1980. “Mathematical Description of the Responses of Simple Cortical Cells*.” *Journal of the Optical Society of America* 70 (11): 1297. https://doi.org/10.1364/JOSA.70.001297.

Markov, Nikola T., Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, et al. 2014. “Anatomy of Hierarchy: Feedforward and Feedback Pathways in Macaque Visual Cortex.” *Journal of Comparative Neurology* 522 (1): 225–59. https://doi.org/10.1002/cne.23458.

Marr, D., S. Ullman, and Sydney Brenner. 1981. “Directional Selectivity and Its Use in Early Visual Processing.” *Proceedings of the Royal Society of London. Series B. Biological Sciences* 211 (1183): 151–80. https://doi.org/10.1098/rspb.1981.0001.

Marr, David. 1982. “Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.” In *New York, NY: W.H. Freeman and Company*. CUMINCAD. http://papers.cumincad.org/cgi-bin/works/Show?fafa.

Mazer, James A., William E. Vinje, Josh McDermott, Peter H. Schiller, and Jack L. Gallant. 2002. “Spatial Frequency and Orientation Tuning Dynamics in Area V1.” *Proceedings of the National Academy of Sciences* 99 (3): 1645–50. https://doi.org/10.1073/pnas.022638499.

McWalter, Richard, and Josh H. McDermott. 2018. “Adaptive and Selective Time Averaging of Auditory Scenes.” *Current Biology* 28 (9): 1405-1418.e10. https://doi.org/10.1016/j.cub.2018.03.049.

Motter, B. C. 2009. “Central V4 Receptive Fields Are Scaled by the V1 Cortical Magnification and Correspond to a Constant-Sized Sampling of the V1 Surface.” *Journal of Neuroscience* 29 (18): 5749–57. https://doi.org/10.1523/JNEUROSCI.4496-08.2009.

Oberfeld, Daniel, and Patricia Stahn. 2012. “Sequential Grouping Modulates the Effect of Non-Simultaneous Masking on Auditory Intensity Resolution.” *PLoS ONE* 7 (10). https://doi.org/10.1371/journal.pone.0048054.

Okazawa, Gouki, Satohiro Tajima, and Hidehiko Komatsu. 2015. “Image Statistics Underlying Natural Texture Selectivity of Neurons in Macaque V4.” *Proceedings of the National Academy of Sciences* 112 (4): E351–60. https://doi.org/10.1073/pnas.1415146112.

Okazawa, Gouki, Satohiro Tajima, and Hidehiko Komats. 2017. “Gradual Development of Visual Texture-Selective Properties Between Macaque Areas V2 and V4.” *Cerebral Cortex* 27 (10): 4867–80. https://doi.org/10.1093/cercor/bhw282.

Olshausen, Bruno A., and David J. Field. 1996. “Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images.” *Nature* 381 (6583): 607–9. https://doi.org/10.1038/381607a0.

Overvliet, K. E., and B. Sayim. 2016. “Perceptual Grouping Determines Haptic Contextual Modulation.” *Vision Research*, Quantitative Approaches in Gestalt Perception, 126 (September): 52–58. https://doi.org/10.1016/j.visres.2015.04.016.

Paradiso, Michael A., and Ken Nakayama. 1991. “Brightness Perception and Filling-In.” *Vision Research* 31 (7): 1221–36. https://doi.org/10.1016/0042-6989(91)90047-9.

Pasupathy, Anitha, Taekjun Kim, and Dina V Popovkina. 2019. “Object Shape and Surface Properties Are Jointly Encoded in Mid-Level Ventral Visual Cortex.” *Current Opinion in Neurobiology*, Computational Neuroscience, 58 (October): 199–208. https://doi.org/10.1016/j.conb.2019.09.009.

Pecka, Michael, Yunyun Han, Elie Sader, and Thomas D. Mrsic-Flogel. 2014. “Experience-Dependent Specialization of Receptive Field Surround for Selective Coding of Natural Scenes.” *Neuron* 84 (2): 457–69. https://doi.org/10.1016/j.neuron.2014.09.010.

Peichl, L., and H. Wässle. 1979. “Size, Scatter and Coverage of Ganglion Cell Receptive Field Centres in the Cat Retina.” *The Journal of Physiology* 291 (1): 117–41. https://doi.org/10.1113/jphysiol.1979.sp012803.

Pelli, D. G., Melanie Palomares, and Najib J. Majaj. 2004. “Crowding Is Unlike Ordinary Masking: Distinguishing Feature Integration from Detection.” *Journal of Vision* 4 (12): 12–12. https://doi.org/10.1167/4.12.12.

Pelli, Denis G., and Katharine A Tillman. 2008. “The Uncrowded Window of Object Recognition.” *Nature Neuroscience* 11 (10): 1129–35. https://doi.org/10.1038/nn1208-1463b.

Peterhans, E., and R. von der Heydt. 1993. “Functional Organization of Area V2 in the Alert Macaque.” *European Journal of Neuroscience* 5 (5): 509–24. https://doi.org/10.1111/j.1460-9568.1993.tb00517.x.

Petrov, Yury, and Olga Meleshkevich. 2011. “Asymmetries and Idiosyncratic Hot Spots in Crowding.” *Vision Research* 51 (10): 1117–23. https://doi.org/10.1016/j.visres.2011.03.001.

Petrov, Yury, Ariella V. Popple, and Suzanne P. McKee. 2007. “Crowding and Surround Suppression: Not to Be Confused.” *Journal of Vision* 7 (2): 12–12. https://doi.org/10.1167/7.2.12.

Polat, Uri, and Dov Sagi. 1994. “The Architecture of Perceptual Spatial Interactions.” *Vision Research* 34 (1): 73–78. https://doi.org/10.1016/0042-6989(94)90258-5.

Poort, Jasper, Matthew W. Self, Bram van Vugt, Hemi Malkki, and Pieter R. Roelfsema. 2016. “Texture Segregation Causes Early Figure Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex.” *Cerebral Cortex* 26 (10): 3964–76. https://doi.org/10.1093/cercor/bhw235.

Portilla, Javier, and Eero P. Simoncelli. 2000. “A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients.” *International Journal of Computer Vision* 40 (1): 49–70. https://doi.org/10.1023/A:1026553619983.

Qiu, Cheng, Daniel Kersten, and Cheryl A. Olman. 2013. “Segmentation Decreases the Magnitude of the Tilt Illusion.” *Journal of Vision* 13 (13): 19–19. https://doi.org/10.1167/13.13.19.

Rao, R P, and D H Ballard. 1999. “Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects.” *Nature Neuroscience* 2 (1): 79–87. https://doi.org/10.1038/4580.

Riesenhuber, Maximilian, and Tomaso Poggio. 1999. “Hierarchical Models of Object Recognition in Cortex.” *Nature Neuroscience* 2 (11): 7.

Roe, Anna W., Leonardo Chelazzi, Charles E. Connor, Bevil R. Conway, Ichiro Fujita, Jack L. Gallant, Haidong Lu, and Wim Vanduffel. 2012. “Toward a

Unified Theory of Visual Area V4.” *Neuron* 74 (1): 12–29. https://doi.org/10.1016/j.neuron.2012.03.011.

Roelfsema, Pieter R. 2006. “Cortical Algorithms for Perceptual Grouping.” *Annual Review of Neuroscience* 29 (1): 203–27. https://doi.org/10.1146/annurev.neuro.29.051605.112939.

Rosenholtz, Ruth. 2014. “Texture Perception.” In *The Oxford Handbook of Perceptual Organization*. https://doi.org/10.1093/oxfordhb/9780199686858.013.058.

Rosenholtz, Ruth.. 2016. “Capabilities and Limitations of Peripheral Vision.” *Annual Review of Vision Science* 2 (1): 437–57. https://doi.org/10.1146/annurev-vision-082114-035733.

Rosenholtz, Ruth, Jie Huang, Alvin Raj, Benjamin J. Balas, and Livia Ilie. 2012. “A Summary Statistic Representation in Peripheral Vision Explains Visual Search.” *Journal of Vision* 12 (4): 14–14. https://doi.org/10.1167/12.4.14.

Rosenholtz, Ruth, Dian Yu, and Shaiyan Keshvari. 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.

Rust, Nicole C., and James J. DiCarlo. 2010. “Selectivity and Tolerance (‘Invariance’) Both Increase as Visual Information Propagates from Cortical Area V4 to IT.” *Journal of Neuroscience* 30 (39): 12978–95. https://doi.org/10.1523/JNEUROSCI.0179-10.2010.

Saarela, Toni P., and Michael H. Herzog. 2009. “Size Tuning and Contextual Modulation of Backward Contrast Masking.” *Journal of Vision* 9 (11): 21–21. https://doi.org/10.1167/9.11.21.

Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. 2017. “Dynamic Routing Between Capsules.” *ArXiv:1710.09829 [Cs]*, November. http://arxiv.org/abs/1710.09829.

Schwartz, Odelia, and Eero P. Simoncelli. 2001. “Natural Signal Statistics and Sensory Gain Control.” *Nature Neuroscience* 4 (8): 819–25. https://doi.org/10.1038/90526.

Seriès, Peggy, Jean Lorenceau, and Yves Frégnac. 2003. “The ‘Silent’ Surround of V1 Receptive Fields: Theory and Experiments.” *Journal of Physiology-Paris*, Neuroscience and Computation, 97 (4): 453–74. https://doi.org/10.1016/j.jphysparis.2004.01.023.

Simoncelli, Eero P, and Bruno A Olshausen. 2001. "Natural Image Statistics and Neural Representation." *Annual Review of Neuroscience* 24 (1): 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193.

Sincich, Lawrence C., and Jonathan C. Horton. 2005. "THE CIRCUITRY OF V1 AND V2: Integration of Color, Form, and Motion." *Annual Review of Neuroscience* 28 (1): 303–26. https://doi.org/10.1146/annurev.neuro.28.061604.135731.

Strasburger, Hans, Ingo Rentschler, and Martin Jüttner. 2011. "Peripheral Vision and Pattern Recognition: A Review." *Journal of Vision* 11 (5): 13. https://doi.org/10.1167/11.5.13.

Stürzel, Frank, and Lothar Spillmann. 2001. "Texture Fading Correlates with Stimulus Salience." *Vision Research* 41 (23): 2969–77. https://doi.org/10.1016/S0042-6989(01)00172-9.

Sun, Hsin-Mei, and Benjamin Balas. 2014. "Face Features and Face Configurations Both Contribute to Visual Crowding." *Attention, Perception & Psychophysics* 77: 508–19. https://doi.org/10.3758/s13414-014-0786-0.

Tesileanu, Tiberiu, Mary M Conte, John J Briguglio, Ann M Hermundstad, Jonathan D Victor, and Vijay Balasubramanian. 2020. "Efficient Coding of Natural Scene Statistics Predicts Discrimination Thresholds for Grayscale Textures." Edited by Stephanie Palmer. *ELife* 9 (August): e54347. https://doi.org/10.7554/eLife.54347.

Thomson, Mitchell G. A., and David H. Foster. 1997. "Role of Second- and Third-Order Statistics in the Discriminability of Natural Images." *JOSA A* 14 (9): 2081–90. https://doi.org/10.1364/JOSAA.14.002081.

Tkacik, G., J. S. Prentice, J. D. Victor, and V. Balasubramanian. 2010. "Local Statistics in Natural Scenes Predict the Saliency of Synthetic Textures." *Proceedings of the National Academy of Sciences* 107 (42): 18149–54. https://doi.org/10.1073/pnas.0914916107.

Tolhurst, David J, and Yoav Tadmor. 2000. "Discrimination of Spectrally Blended Natural Images: Optimisation of the Human Visual System for Encoding

Natural Images." *Perception* 29 (9): 1087–1100. https://doi.org/10.1068/p3015.

Tootell, Rb, E Switkes, Ms Silverman, and Sl Hamilton. 1988. "Functional Anatomy of Macaque Striate Cortex. II. Retinotopic Organization." *The Journal of Neuroscience* 8 (5): 1531–68. https://doi.org/10.1523/JNEUROSCI.08-05-01531.1988.

Treisman, Anne. 1996. "The Binding Problem." *Current Opinion in Neurobiology* 6 (2): 171–78. https://doi.org/10.1016/S0959-4388(96)80070-5.

Van Essen, David C., William T. Newsome, and John H. R. Maunsell. 1984. "The Visual Field Representation in Striate Cortex of the Macaque Monkey: Asymmetries, Anisotropies, and Individual Variability." *Vision Research* 24 (5): 429–48. https://doi.org/10.1016/0042-6989(84)90041-5.

Vergeer, Mark L T, and Rob van Lier. 2007. "Grouping Effects in Flash-Induced Perceptual Fading." *Perception* 36 (7): 1036–42. https://doi.org/10.1068/p5607.

Vinje, William E., and Jack L. Gallant. 2000. "Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision." *Science* 287 (5456): 1273–76. https://doi.org/10.1126/science.287.5456.1273.

Vinje, William E., and Jack L. Gallant. 2002. "Natural Stimulation of the Nonclassical Receptive Field Increases Information Transmission Efficiency in V1." *Journal of Neuroscience* 22 (7): 2904–15. https://doi.org/10.1523/JNEUROSCI.22-07-02904.2002.

Wagemans, Johan, James Elder, Michael Kubovy, Stephen Palmer, Mary Peterson, Manish Singh, and Rüdiger von der Heydt. 2012. "A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure–Ground Organization." *Psychological Bulletin* 138 (6): 1172–1217. https://doi.org/10.1037/a0029333.

Wallace, Julian M, and Bosco S Tjan. 2011. "Object Crowding." *Journal of Vision* 11 (6): 19. https://doi.org/10.1167/11.6.19.

Wallis, Thomas S. A., Christina M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Matthias Bethge. 2019. "Image Content Is More Important than Bouma's Law for Scene Metamers." *ELife* 8: e42512. https://doi.org/10.7554/eLife.42512.

Wallisch, Pascal, and J. Anthony Movshon. 2008. “Structure and Function Come Unglued in the Visual Cortex.” *Neuron* 60 (2): 195–97. https://doi.org/10.1016/j.neuron.2008.10.008.

Wandell, Brian A. 1995. *Foundations of Vision*. Sinauer Associates. insights.ovid.com.

Whitney, David, and Dennis M. Levi. 2011. “Visual Crowding: A Fundamental Limit on Conscious Perception and Object Recognition.” *Trends in Cognitive Sciences* 15 (4): 160–68. https://doi.org/10.1016/j.tics.2011.02.005.

Whitney, David, and Allison Yamanashi Leib. 2018. “Ensemble Perception.” *Annual Review of Psychology* 69 (1): 105–29. https://doi.org/10.1146/annurev-psych-010416-044232.

Willenbockel, Verena, Javid Sadr, Daniel Fiset, Greg O. Horne, Frédéric Gosselin, and James W. Tanaka. 2010. “Controlling Low-Level Image Properties: The SHINE Toolbox.” *Behavior Research Methods* 42 (3): 671–84. https://doi.org/10.3758/BRM.42.3.671.

Xing, Jing, and David J Heeger. 2000. “Center-Surround Interactions in Foveal and Peripheral Vision.” *Vision Research* 40 (22): 3065–72. https://doi.org/10.1016/S0042-6989(00)00152-8.

Young, Richard A. 1985. “THE GAUSSIAN DERIVATIVE MODEL FOR MACHINE AND BIOLOGICAL IMAGE PROCESSING.” In *Proceedings of the Conference of the Society of Photographic Scientists and Engineers*, 64–70. Springfield: SPSE.

Young, Richard A. 1986. “THE GAUSSIAN DERIVATIVE MODEL FOR MACHINE VISION: VISUAL CORTEX SIMULATION.” 5323.

Young, Richard, Ronald Lesperance, and W. Weston Meyer. 2001. “The Gaussian Derivative Model for Spatial-Temporal Vision: I. Cortical Model.” *Spatial Vision* 14 (3–4): 261–319. https://doi.org/10.1163/156856801753253582.

Zavitz, Elizabeth, and Curtis L Baker. 2014. “Higher Order Image Structure Enables Boundary Segmentation in the Absence of Luminance or Contrast Cues.” *Journal of Vision* 14 (2014): 14. https://doi.org/10.1167/14.4.14.

Zetzsche, C., and F. Rhrbein. 2001. “Nonlinear and Extra-Classical Receptive Field Properties and the Statistics of Natural Scenes.” *Network: Computation in Neural Systems* 12 (3): 331–50. https://doi.org/10.1088/0954-898X/12/3/306.

Zetzsche, Christoph, and Ulrich Nuding. 2005. “Nonlinear and Higher-Order Approaches to the Encoding of Natural Scenes.” *Network: Computation in Neural Systems* 16 (2–3): 191–221. https://doi.org/10.1080/09548980500463982.

Zhaoping, L. 1998. “A Neural Model of Contour Integration in the Primary Visual Cortex.” *Neural Computation* 10: 903–40.

Ziemba, Corey M., Jeremy Freeman, J. Anthony Movshon, and Eero P. Simoncelli. 2016. “Selectivity and Tolerance for Visual Texture in Macaque V2.” *Proceedings of the National Academy of Sciences* 113 (22): E3140–49. https://doi.org/10.1073/pnas.1510847113.