Putting low-level vision into global context: Why vision cannot be reduced to basic circuits

Michael H. Herzog a,⇑, Evelina Thunell a, Haluk Ögmen b

a Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
b Department of Electrical and Computer Engineering, Center for Neuro-Engineering and Cognitive Science, University of Houston, TX, USA

ARTICLE INFO
Article history:
Received 14 February 2015
Received in revised form 28 July 2015
Accepted 18 September 2015
Available online 17 October 2015

Keywords:
Crowding
Non-retinotopic processing
Grouping
Pooling models

ABSTRACT
To cope with the complexity of vision, most models in neuroscience and computer vision are of hierarchical and feedforward nature. Low-level vision, such as edge and motion detection, is explained by basic low-level neural circuits, whose outputs serve as building blocks for more complex circuits computing higher level features such as shape and entire objects. There is an isomorphism between states of the outer world, neural circuits, and perception, inspired by the positivistic philosophy of the mind. Here, we show that although such an approach is conceptually and mathematically appealing, it fails to explain many phenomena including crowding, visual masking, and non-retinotopic processing.

1. Introduction

Understanding a visual scene usually requires taking information across the entire visual field into account. Most models of vision try to break down the complexity of vision by a hierarchical, feedforward filtering approach. Based on the findings by Hubel and Wiesel (1959), neurons in the primary visual cortex (V1) extract lines by combining inputs from LGN neurons as a first step. The outputs of these neurons serve as building blocks for the next step: V1 neurons project to V2 neurons and so on. With each step, receptive field sizes increase and information is pooled across larger and larger regions of the visual field, eventually allowing higher cortical neurons to code for high-level features such as shapes and objects (see for example DiCarlo, Zoccolan, & Rust, 2012; Riesenhuber & Poggio, 1999; Thorpe, Fize, & Marlot, 1996). In a cartoon version, a “square neuron” receives input from neurons sensitive to the vertical and horizontal lines that make up the square.

At each stage, neurons are thought to process stimuli in a highly stereotyped fashion. For example, the response of a V1 neuron to a vertical line is independent of whether the line is presented alone or as part of a square because, to reduce complexity and keep processing simple, there are neither lateral nor top-down interactions (except for very local ones as shown in Fig. 1). Therefore, the neuron is “blind” to the horizontal lines of the square and to the overall shape. Low-level determines high-level analysis, but not the other way around. This highly stereotypical, hierarchical and feedforward processing is mathematically appealing, breaking down a seemingly impossible problem into treatable subproblems. Accordingly, research of the last 50 years has focused on understanding the detailed characteristics of low-level visual processing. The subjective aspects of perception, such as perceptual grouping and Gestalt, are assumed to emerge naturally in the subsequent stages, combining the outputs of low-level processing.

Within this framework, there is an important implicit assumption: There is an isomorphism between external world states, their neural presentations, and the corresponding percepts. For example in surround suppression, visibility of a grating patch increases monotonically with the patch size up to a certain point, beyond which visibility decreases. How can this non-monotonic dependency of performance on patch size be explained? Fig. 1 shows three tentative circuits. The circuits differ mainly in at which stage inhibition comes into play. The neurometric function matches the psychometric function, and because of this isomorphism, subjective terms can be eliminated. In this sense, vision research is well in the tradition of philosophical eliminativism (see for example Churchland, 1981).

Similar circuits are omnipresent in all fields of low-level vision, for example in crowding. In crowding, performance on a target strongly deteriorates when adding elements next to the target (Fig. 2). Crowding has been explained by pooling models, where
signals of neurons with smaller receptive fields are integrated by neurons with larger receptive fields, causing target irrelevant information from the flankers to be averaged with the target signal (Fig. 2; Pelli, Palomares, & Majaj, 2004; Wilkinson, Wilson, & Ellemberg, 1997). These are examples of classic hierarchical and feedforward models, with pooling as the mechanism for combining information. In this sense, crowding is a reflection of the unavoidable limitations of the visual system. The brain simply cannot do better because it needs to pool information to allow for object recognition. As mentioned, similar circuits have been proposed in all fields of low-level visual processing. Importantly, changes in the firing of an output neuron are isomorphic to changes in perception. Reproduced from Smith (2006).

Fig. 1. Three possible mechanisms for surround suppression (Smith, 2006). An output neuron (gray triangle) is activated by the central part of the grating, which falls in the classical receptive field of the neuron. The neuron is inhibited by the surround. The white triangles represent excitatory neurons and the black disks represent inhibitory neurons. (A) Lateral connection model. V1 neurons with receptive fields centered on the surround suppress the central neuron via lateral inhibition. (B) An extrastriate neuron receives input from many V1 neurons, and then inhibits the central neuron via an inhibitory interneuron. (C) Surround suppression is generated within LGN. Similar circuits have been proposed for all types of low-level visual processing. Importantly, changes in the firing of an output neuron are isomorphic to changes in perception. Reproduced from Smith (2006).

Fig. 2. (A) Observers indicated whether a vernier was offset to the left or to the right (the vernier is shown next to the dashed line). We determined the offset size for which 75% correct responses were obtained (threshold). Results for the flanked conditions are plotted in terms of threshold elevation compared to this unflanked vernier condition (dashed line), i.e., thresholds divided by the threshold of the unflanked condition. A threshold “elevation” of 1.0 indicates no crowding; values larger than 1.0 indicate crowding. Vernier offset discrimination deteriorates when a single line is added on each side (a). When the vernier is flanked by rectangles, performance improves even though the flankers of (a) are part of the rectangles. Adapted from Sayim et al. (2010) and Manassi et al. (2012). (B) In classic models of object recognition, a visual stimulus is analyzed by sets of filter banks. In the first step of filtering, lines and edges are extracted. In the subsequent step, neurons coding for nearby elements combine the outputs of the lower level neurons. Usually, pooling is thought to occur only for neurons coding for similar features, for example vertical but not horizontal lines, in agreement with most psychophysical results (e.g., Kooi et al., 1994). For this reason, the configuration with the vernier plus flanking lines gives a very similar output of filtering as the vernier plus boxes configuration. However, such pooling cannot explain the large effects of configuration shown in A. (C) We propose that our results can only be explained by a flexible grouping stage that determines which elements interfere with each other (by whatever mechanism).
In addition to being stereotypical, hierarchical, and feedforward, low-level feature coding is often thought to be retinotopic. Neighboring elements in the visual scene activate neighboring neurons in the retina and early visual areas. The position of a neuron in cortex is often assumed to code for elements in a certain position in the visual scene. For example, in visual search a retinotopic master map has been proposed to operate on retinotopic feature maps, where the location in the map codes for the position of the element (Treisman & Gelade, 1980).

In some way, it is an appealing idea that the human brain solves the complex problems of vision by breaking down processing into basic circuits, which operate in retinotopic coordinates more or less independently of other low- and high-level circuits. Recurrent processing across the entire visual field is hardly mathematically treatable. Still, as we will show here, low-level vision is not stereotypical, hierarchical, feedforward, and retinotopic, but a complex recurrent process in which the neural representations of elements across large parts of the visual field interact. More importantly, our results question the idea of a simple isomorphism between outer world states, basic circuits, and low-level percepts, as well as classical hierarchical and feedforward models.

We propose that models of visual processing need to take into account a “flexible” grouping stage, which operates on all levels of visual processing (Fig. 2). Vision cannot easily be broken down into basic sub-processes that filter information in a highly stereotypical fashion within small regions of the visual field. Grouping can depend on small spatial changes, “invisible” to low-level processes, which lead to strong changes of perception. Because of the complex grouping process, there cannot be an isomorphism between external world states, basic stereotyped circuits, and perception. Subjective terms such as grouping can, at least at the moment, not be eliminated.

2. Review of results

2.1. How grouping determines vernier acuity in crowding

We presented a vernier (two vertical lines with a small horizontal offset either to the left or right). Observers indicated the offset direction (Fig. 2A, dashed line). In the next condition, we added one vertical line to the left and one to the right of the vernier, inducing a classic crowding effect: performance strongly deteriorated (Fig. 2A, condition a). We then added additional lines, turning the flanking lines into parts of boxes. Performance strongly improved compared to the previous condition (Fig. 2A, condition b; Manassi, Sayim, & Herzog, 2012; Sayim, Westheimer, & Herzog, 2010). Most models of crowding cannot easily explain these results because the flanking lines next to the vernier are the very same in both conditions. Hence, any model that relies on stereotypical filtering should lead to the same performance level (Fig. 2B). However, the performance levels differ. We propose that it is the good Gestalt of the boxes that releases the target from crowding: Because the lines belong to the boxes, they no longer group with the vernier and accordingly crowding is reduced. In other words, grouping determines crowding. These results hold true for both foveal (Sayim et al., 2010) and peripheral crowding (Manassi et al., 2012). Similar crowding results were found in haptics (Overvliet & Sayim, 2015), in audition (Oberfeld, Stahn, & Kuta, 2014), and in visual crowding with other stimuli than verniers (Chakravarthi & Pelli, 2011; Joo, Boynton, & Murray, 2012; Levi & Carney, 2009; Livne & Sagi, 2007; Livne & Sagi, 2010; Louie, Bressler, & Whitney, 2007; Pöder, 2006; Robol, Casco, & Dakin, 2012; Saarela, Sayim, Westheimer, & Herzog, 2009; Woldorf & Chambers, 1983; Yeotikar, Khuu, Asper, & Suttle, 2011).

As mentioned, the aim of classic models of vision is to directly link perception to neural mechanisms such as pooling and thus naturalize and eliminate the subjectivity of perception. In this view, basic feature analysis is independent of the global stimulus configuration simply because it precedes the processing of global configurations. However, it seems that contextual modulation is very little determined by low-level feature interactions, and instead depends heavily on the overall stimulus configuration. Thus, vision cannot be explained by a stereotyped filtering process. For the same reason, vision cannot be reduced to basic circuits independent of top-down processing.

It is important to note that grouping does not explain why there is crowding, i.e., why elements interfere with each other in the first place. Grouping only determines which elements interfere. It may
well be that once grouping is established, pooling (or some similar mechanism) comes into play. However, this provokes the question of why pooling deteriorates performance if the human brain is in principle able to resolve fine-grained stimulus details in crowded conditions. Classic models propose that crowding occurs because of a low-level bottleneck when elements are presented in clutter. We suggested that there is no low-level bottleneck at all (Herzog, Hermens, & Ögmen, 2014).

2.2. How grouping determines Gabor detection in overlay masking

Our results are not restricted to vernier stimuli and crowding. Saarela and Herzog (2009) measured contrast detection thresholds for a Gabor target masked by a small grating patch. The mask was presented immediately after the target at the same spatial location (backward masking). A small backward mask, matched to the target in orientation, phase, and spatial frequency, had a large masking effect (Fig. 3a and b). This mask was then embedded in various surrounds. When embedded in a contiguous surround matched in orientation, phase, and spatial frequency, masking was greatly reduced (i.e., thresholds improved; condition c in Fig. 3). Masking was unaffected by the surround, however, when the surround was segregated from the central mask by a small gap (condition d). The same was true no matter how the center and surround differed: by orientation, spatial frequency, or spatial phase (remaining conditions e-l in Fig. 3).

We argue that the gaps influence target processing because they segregate the surround from the target. As in crowding, how the elements segregate, i.e., ungroup, determines how they interact with each other. We suggest that similar mechanisms are at play in surround suppression (Fig. 1), which is very similar to overlay masking. Similarly, in the tilt illusion, illusion strength reduces when target and flankers segregate from each other (Qiu, Kersten, & Olman, 2013).

2.3. Grouping across large parts of the visual field determines crowding

We presented a square at 9° of eccentricity to the right of fixation, and asked observers to discriminate whether the x-axis of the square was longer or shorter than the y-axis. Performance strongly deteriorated when the central square was flanked by three additional squares on each side (a classic crowding effect; Fig. 4A). Next, we presented a vernier and asked observers to discriminate its offset. When a single square surrounded the vernier, performance strongly deteriorated (Fig. 4B). This is another crowding effect. In the third step, we combined the two conditions. Since the vernier is now surrounded by even more elements, simple pooling models would predict enhanced crowding (Fig. 4B). However, we found the opposite: the crowding of the vernier was almost completely abolished. Crowding of crowding thus led to uncrowding (results from Manassi, Sayim, & Herzog, 2013). This reduction of crowding increases smoothly with the number of squares (Fig. 4C). The configuration with the 7 squares covers up to 17° of the right visual field, i.e., the outmost squares are well beyond Bouma’s window.

As we have seen, crowding depends on grouping of elements over large regions of the visual field. The results cannot be explained by low-level effects but only by shape–shape interactions. This becomes even more obvious when removing the horizontal lines of the contextual squares: the strong crowding is once again “restored” (Fig. 4D). Likewise, rotating the contextual squares leads to strong crowding (Fig. 4D). Here, we have shown a few examples of how the overall configuration of the entire stimulus determined crowding. More can be found in these reviews (Herzog & Manassi, 2015; Herzog, Sayim, Chicherov, & Manassi, 2015).

![Fig. 4.](image-url)

(A) Stimuli were presented at 9° of eccentricity. Observers were asked to discriminate whether a rectangle was wider or narrower along the horizontal axis (“x”, dashed line). We determined the threshold width for which 75% correct responses were obtained. When the rectangle was flanked by three squares on each side, thresholds strongly increased compared to when it was presented alone. This is a classic crowding effect. (B) Observers were asked to discriminate the vernier offset direction (left vs. right). We determined the offset size for which 75% correct responses occurred (dashed line). Thresholds increased when the vernier was embedded in a square. This is another crowding effect. Surprisingly, thresholds decreased and crowding almost vanished when we combined the two conditions. Hence, crowding of crowding leads to uncrowding. (C) Performance increases gradually with the number of flanking squares. In the 7 squares condition, the stimulus spans up to 17° of the right visual field, showing that even elements far away from the target can have an uncrowding effect. (D) Removing the horizontal lines from the contextual squares leads to strong crowding. Hence, the improved performance cannot be explained by the additional vertical lines alone. Rotating the contextual squares also leads to strong crowding, showing that shape–shape interactions determine crowding. From: Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. Journal of Vision, 13, 10–10. Copyright: Association for Research in Vision and Ophthalmology (ARVO ©).
Fig. 5. (A) Feature fusion. A right or left offset vernier is followed by a second vernier with the opposite offset direction (anti-vernier, AV). Only one vernier is perceived. Observers cannot tell whether the first or second vernier is offset to the right (results not shown; Scharnowski et al., 2009). (B) When the second vernier is flanked by 12 copies of itself on each side, feature fusion ceases almost completely. The first vernier becomes visible as an element in its own right, appearing to be superimposed on the grating. (C) Vernier dominance. In each trial, we determined whether observers’ responses matched the offset direction of the first vernier. Hence, a 100% score means that observers always responded according to the first vernier, 0% always according to the second vernier, and 50% that both verniers contributed equally on average. Performance is at ceiling for a single vernier. Adding the anti-vernier leads to anti-vernier dominance (in accordance with the fact that in fusion, the latter element dominates; Efron, 1967). When there are 12 copies of the second vernier on either side of it, the first vernier becomes visible and performance is almost as good as for the single vernier condition (only 2 * 6 verniers are shown here). In the center, the very same vernier is presented as in the previous condition. It seems that spatial integration trumps temporal fusion. Adapted from Hermens et al. (2009).

2.4. Grouping on the short-term scale

In the previous experiment, we showed that contextual modulation depends on grouping over large parts of the visual field. Here, we show that grouping operates similarly on fine-grained spatio-temporal scales. When a right offset vernier is followed by a left offset vernier in immediate succession (or the other way around; Fig. 5A), an almost aligned vernier is perceived (Herzog, Parish, Koch, & Fahle, 2003). Presentation times can be as short as 10 ms for each vernier. This outcome is typically explained by integration masking where the two vernier offsets are invisible because of a lack of temporal resolution (Scheerer, 1973; Turvey, 1973). The human brain is simply not able to render the two elements individually visible. However, this is not true. When we exchanged the vernier in the second frame for an array of 25 verniers, the first vernier became visible and discrimination of its offset was only slightly deteriorated compared to when it was presented alone (Fig. 5B and C; Hermens, Scharnowski, & Herzog, 2009). We propose that since the 25 verniers are grouped into an array of elements independently of the first vernier, both the first vernier and the array of verniers are rendered visible as objects in their own right (Herzog & Fahle, 2002; Herzog & Koch, 2001). Hence, the human brain can easily resolve the two verniers and their offsets, when they are not grouped together. We proposed that spatial grouping of the vernier array undoes the temporal integration of the first and the second vernier, even though both are presented at the same retinal location.

2.5. Non-retinotopic processing

Here, we show that there is a flexible grouping stage also in non-retinotopic processing, where features are integrated across time and space. In the sequential metatcontrast masking paradigm, a central vernier offset to the left or right is followed by line elements without offsets. A cue indicates whether observers have to attend to the left or right stream (right in this example). Although the central vernier is invisible, observers can report its offset, because it is perceived in the attended stream. When the second element of the attended stream has an offset in the opposite direction compared to the central vernier, the two offsets cancel out and performance drops to about 50%. When the grouping of the central vernier to both streams is disambiguated by presenting an element to the right of the vernier, the vernier no longer groups with the right stream and observers report predominantly the offset of the second element in the stream (performance is below 50%). Interestingly, the non-retinotopic, spatio-temporal grouping here overrides the retinotopic overlap of the vernier and the second element of the stream. If the line in the first frame is presented to the left, the vernier again groups with the right stream and integrates with the second line offset. Performance is accordingly at the 50% level. From: Otto, T. U., Ogmen, H., & Herzog, M. H. (2006). The flight path of the phoenix–The visible trace of invisible elements in human vision. Journal of Vision, 6, 1079–1086. Copyright: Association for Research in Vision and Ophthalmology (ARVO ©).
elements are grouped across space and time. In the second step, split up in two stages. First, the visual system determines which ISIs (Fig. 7C), again revealing the horizontal and vertical motions. The group motion can also be disrupted by omitting the dots. Now perceived to move horizontally and vertically. The motion percept breaks down and the disks integrate retinotopically. Adaptation and visual backward masking, on the other hand, occur in retinotopic coordinates (Boi, Ögmen, & Herzog, 2011; Noory, Herzog, & Ögmen, 2015, but see Lin & He, 2012).

Non-retinotopic integration has been found also for other features. For example, Shimozaki and colleagues (Shimozaki, Eckstein, & Thomas, 1999) showed that the luminance of an object influences succeeding luminance judgments after the object moved. Also color can integrate non-retinotopically: when the color of moving bars alternates between red and green, observers perceive the mixed color (yellow) even though the red and green colors never appear in the same retinal location (Nishida, Watanabe, Kuriki, & Tokimoto, 2007). Further, Kawabe and colleagues (Kawabe, 2008) showed that the perceived size of a disk is biased towards the size of a preceding disk appearing in a different location. When two possible apparent motion streams were presented, a bias according to the attended stream was found.

Non-retinotopic integration can also be induced using the Ternus–Pikler display. In the main condition of this paradigm, three disks are presented alternatingly to the left and to the right (Fig. 7A), interleaved by blank pauses (Inter-Stimulus Intervals; ISIs). The disks are perceived to move back and forth as a group. The apparent motion creates a non-retinotopic reference frame (as indicated by the arrows), in which the white dots in the center disks are integrated across frames. Because the horizontal motion of the disks is subtracted from the dot motion, clockwise rotation is perceived. When the outer disks are removed (Fig. 7B), the group motion percept breaks down and the disks integrate retinotopically. The dots are now perceived to move horizontally and vertically. The group motion can also be disrupted by omitting the ISIs (Fig. 7C), again revealing the horizontal and vertical motions.

We propose that non-retinotopic feature integration can be split up in two stages. First, the visual system determines which elements are grouped across space and time. In the second step, the features of the grouped elements integrate with each other (Clarke, Ögmen, & Herzog, 2016, but see Clarke, Repnow, Ögmen, & Herzog, 2013 and Pooresmaeili, Cicchini, Morrone, & Burr, 2012).

Non-retinotopic feature integration can be

2.6. Grouping and the Gestalt laws

It seems that the overall stimulus configuration determines crowding, pointing to the role of Gestalt factors. However, the following experiment shows that crowding cannot easily be explained by simple Gestalt rules (Saarela, Westheimer, & Herzog, 2010). Observers identified the orientation of a peripheral target “T” when it was flanked by other “T’s” in various configurations. When the spacing was regular and relatively tight, crowding was strong (Fig. 8, “Tight” condition). When the flanks closest to the target were moved away, crowding was reduced, as expected (“Shifted”). When the remaining flanks were also moved away, so that the spacing throughout the stimulus was regular, crowding increased (“Wide”). Finally, a condition having even more flanks led to weaker crowding (“Added”).

The same pattern of results was found with Gabor patches (Saarela et al., 2010). The effectiveness of the flanks coincided with the subjective reports of the observers: when the target perceptually appeared to belong to the same “whole” as the flanks, crowding was strong. Conversely, when the target appeared to stand out, crowding was weaker. These results are in line with earlier studies that correlated observers’ subjective ratings of target conspicuity with other stimuli (Malania, Herzog, & Westheimer, 2007; Manassi et al., 2012; Saarela et al., 2009). The bottom line is that what determines the strength of crowding is how the target is grouped with the flanks.

It has been established that grouping cannot easily be predicted by the well known Gestalt laws. This is also true in our data.
For example, in Fig. 8, it is unclear why spacing regularity trumps proximity, which is usually considered one of the strongest grouping cues. It seems that what matters for crowding is the global stimulus configuration (see also Dakin, Greenwood, Carlson, & Bex, 2011; Maus, Fischer, & Whitney, 2011; Wallis & Bex, 2011). The appearance of the whole determines performance of its parts, in analogy to the famous quote by Wertheimer: “the whole determines the appearance of its parts” (Wertheimer, 1923).

3. Discussion

A hundred years ago, the Gestaltists came up with the first coherent research program of perception. In the spirit of the emerging science of psychology, the subjective aspects of perception were key, particularly, how the perception of parts gives rise to the perception of wholes. Basic rules were proposed, known as the Gestalt laws, including the grouping keys of proximity, similarity, and Prägnanz. The rules were mainly derived from reports of how people subjectively grouped elements (for example Wagemans et al., 2012; Wertheimer, 1923). At the same time, the Wiener Kreis and other positivistic philosophers tried to overcome the highly subjective philosophy of rationalism (Fichte, 1794; Hegel, 1807), which dominated continental philosophy for half a century, by a rigorous program based on basic observations (“sense data”), such as “there is a vertical line at a certain position in space and time”. These basic percepts were thought to be undeniable true and the perfect basis for all sciences. In most of the theories, there is an isomorphism between outer world properties, mental states, and perception (see for example Wittgenstein, 1922). In a contemporary neuroscience interpretation of Wittgenstein’s position, a fork to the left of a knife in the external world is represented by a “fork” and a “knife” neuron, with the fork neuron located next to the knife neuron, i.e., the encoding is retinotopic. Whereas the Gestaltists tried to explain subjectivity by determining the basic rules, philosophical positivism was aimed at eliminating subjectivity by objective terms (Churchland, 1981). The Gestalt program flourished for about 20 years and then more or less disappeared, while positivism has implicitly or explicitly become the backbone of today’s neuroscience, computer vision, and the psychology of perception.

3.1. Neural circuits and perception

There are four fundamental aspects of classic models of vision. First, to cope with the complexity of vision, it is assumed that visual analysis can be broken down into basic, independent processing steps, where each step is mathematically well treatable. Low-level analysis is performed independently for small regions of the visual field and is highly stereotypic. For example, line extraction is largely independent from other low- and high-level processes. Low-level processing determines high-level processing, and not the other way around.

Second, low-level processing occurs mainly in retinotopic coordinates. The location of a neuron codes for the location of the stimulus in the visual field.

Third, there is an isomorphism between external world states, neural responses, and perception. In most models of low-level vision, there is an output neuron within a basic circuit, which determines perception in a similar way as in Fig. 1. The basic circuits in the models are designed in such a way that the neurometric function of the output neuron matches the psychometric function. Research has focused on characterizing these circuits for the last 50 years. The main question was usually at which stage pooling, inhibition, normalization, gain control, or rectification come into play. Circuits as shown in Fig. 1 are ubiquitous in vision research and thought to explain crowding, surround modulation, metacontrast masking, the dipper function in contrast perception, the aperture problem in motion discrimination, all sorts of illusions, and many other phenomena.

Fourth, combining the outputs of basic circuits in a hierarchical and feedforward fashion eventually results in the recognition of objects and scenes. High-level processing is fully determined by low-level processing- and thus the complexity of vision can be broken down into independent, simpler subprocesses. For this reason, top-down projections from higher to lower level areas are thought to be related mainly to attention, consciousness and other modulatory aspects of vision (Lamme, 2006).

Our results challenge these four main characteristics. Of course, because of distortions by blood vessels and the highly different packing and types of photoreceptors in foveal and peripheral vision, there is no one-to-one mapping between external world states and the retinal image. Hence, direct realism cannot be true. Post-retinal processing is often thought to counteract these distortions, restoring the “true” image (indirect realism, sense data theory). Here, we propose that indirect realism cannot be achieved even in low-level stages of visual processing. Vision cannot easily be broken down into basic, stereotyped circuits. Low-level feature processing depends on high-level features as much as the other way around. For example, the high-level Gestalt of the rectangles in Fig. 4 determines vernier acuity, a hallmark of low-level vision.
In the crowding experiments of Fig. 2, the average offset size is about 3 arcmin for the single vernier. Still, squares presented 8.5° away, i.e., 170 times the vernier offset size, can strongly influence vernier acuity. Hence, low-level visual acuity depends on elements across large parts of the visual field and on the overall configuration. Low-level circuits cannot explain these results because first, they are confined to small regions of the visual field, and second, they are “blind” to higher level Gestalt aspects. Likewise, in the Ternus–Pikler display, the global group motion of the disks determines how the motion of the small dot is perceived (Fig. 7). From a retinotopic point of view, the group motion produces a highly non-veridical dot motion percept. Clearly, high-level processing does more than only modulate low-level vision. It seems that the appearance of the whole, i.e., the output of figural processing, determines performance on the constituting elements, similarly to Wertheimer’s conclusion that “the whole determines the appearance of its parts” (Wertheimer, 1923). For the same reasons, our results question the idea of an isomorphic relationship between external world states, basic circuits, and perception. Since everything appears to depend on everything, vision cannot be explained by simple, basic circuits. The output of a neuron sensitive to vernier offsets seems to depend on highly complex processing including low- and high-level interactions.

Finally, our results show that low-level encoding is often not retinotopic, and hardly explained by models which are both hierarchical and feedforward. Hence, at least one of these assumptions needs to be dismissed.

3.2. Grouping and ill-posed problems

We propose that grouping is key in low-level vision. It seems that mutual interference between elements occurs only within groups. For example, only if there is grouping, there is crowding. Grouping occurs on many and across many levels. It can for example occur across large parts of the visual field, and operate on very short time scales, as the feature fusion experiments have shown. Temporal integration of verniers, each presented for only 10 ms, can be undone by contextual verniers (Fig. 6). Because grouping is subjective, we believe that subjective terms cannot easily be eliminated from vision science. In particular, subjective aspects cannot be replaced by models comprising a nested set of simple, stereotypic sub-processes (see also Herzog et al., 2014; Herzog et al., 2015; Herzog & Manassi, 2015). Grouping is flexible in the sense that small changes in the stimuli can drastically change performance. For example, adding one single line in Fig. 6 strongly changes the grouping and accordingly feature integration. Whereas we suggest that the terms grouping and good Gestalt cannot easily be replaced, we propose that basic Gestalt laws, such as proximity and similarity, fail to explain our results for the same reasons as classic vision models fail. The rules are too simple and it remains unclear how they combine with each other (see Saarela et al., 2010, Fig. 8; but see for example also Elder & Goldberg, 2002 and Kubovy, Holcombe, & Wagemans, 1998).

Why are computations and grouping across the entire visual field crucial for perception? The visual system needs to detect for example shadows. This computation includes determining the reflectance of an object, the origin of the illuminance, and the position of the occluder—and that it is an occluder. This can hardly be achieved on a local circuit level (see for example Adelson's shadow illusion Adelson, 1993, but see also Blakeslee & McCourt, 1999). Rather, large-scale integration and figure-ground segmentation are needed. In general, the light that enters a photoreceptor is the product of the light shining on the object (illuminance) and its material properties: luminance = illuminance * reflectance. Hence, the luminance encoded by a photoreceptor does not reveal the material properties of an object, e.g., photoreceptors cannot tell whether or not a banana is yellow. Actually, depending on the light conditions, a banana might “appear” blue to a photoreceptor even if it is just the right yellow. The same is true for spatial properties. There are infinitely many objects that give rise to the very same retinal image. For example, there are infinitely many pens presented at different distances and of different sizes, whose projections on the retina are identical (a small pen close to the eye or a large one further away). In general, the information arriving at the retina is 2D, and the visual system needs to reconstruct the 3D object by taking information across large parts of the visual field into account.

3.3. Low-level bottlenecks and perception

Deteriorated performance in crowding, backward masking, and many other paradigms, is often thought to be caused by limitations and bottlenecks of the visual system (Whitney & Levi, 2011). The idiosyncrasies of basic circuits lead to abnormal perception under certain conditions, and the system can simply not do better. For example, pooling is necessary for object recognition and crowding is just the price to pay. The same is thought to hold true for perceptual distortions in visual illusions. We propose that deteriorated performance and distortions often do not reflect limited low-level vision, but interpretation. For example, in the fusion experiment (Fig. 5), the vernier and the anti-vernier do not fuse because of limitations of the visual system. Rather, the verniers lose their individual visibility because the brain integrates them in a purposeful manner (Herzog et al., 2014). This becomes evident when adding the contextual verniers: The very same vernier and anti-vernier are now treated as two separate elements. Resolution can thus easily be restored. We suggest that the brain renders the verniers invisible in order to perceive one whole element instead of its parts. Likewise, strong crowding of a vernier can easily be undone by adding flanking elements.

We can only speculate about the neural mechanisms underlying the effects of contextual modulation and grouping across the entire visual field. Clearly, simple circuits cannot explain our results. Elements presented far outside the classical receptive field can strongly influence responses of a neuron, often reflected in the notion of the non-classical receptive field. Hence, it may well be that complex figurual processing, as it occurs in crowding and surround suppression, feeds back to a single neuron and changes its firing rate. Our EEG studies suggest that the effects of (un) crowding do not occur before 180 ms, indicating that, indeed, figural processing needs a substantial time to occur (Chicherov, Plomp, & Herzog, 2014; Chicherov & Herzog, 2015). The early EEG responses (the P1 component) do not reflect any figurual processing. Similarly, the transient neural responses to a stimulus do not distinguish between whether or not an element is part of a figure (Lamme, 1995; Zipser, Lamme, & Schiller, 1996). The signal is in the sustained response, i.e., after recurrent processing has had time to occur and grouping across the entire visual field may come into play (Roelfsema, Lamme, & Spekreijse, 1998).

Science progresses by falsification. In this sense, excluding hierarchical, feedforward models may be seen as progress. However, these models make up only a very small fraction of the potential models which include the large set of recurrent models (see for example Clarke, Herzog, & Francis, 2014). To make the situation even further complicated, recurrent models are a mathematical challenge since small changes in the parameters can lead to qualitatively very different behaviors, such as stable fix points, oscillations, and chaos. This is one reason why the classic approach is so widespread, and it is with reluctance that we dismiss it in favor of more complicated models. Likewise, the isomorphic approach, aimed at eliminating the subjective aspects of perception, is appealing as is the idea of stereotypic filtering. It is
again reluctantly that we note that we cannot reduce vision to basic circuits that determine the basics of perception.

Acknowledgments

This work was supported by the Swiss National Science Foundation (SNF) project “Basics of visual processing: from retinotopic encoding to non-retinotopic representations”. We would like to thank Toni Saarela and Mauro Manassi for useful comments on the manuscript.

References


