frontiers in
COMPUTATIONAL NEUROSCIENCE

HYPOTHESIS AND THEORY ARTICLE
published: 12 December 2014
doi: 10.3389/fncom.2014.00158

A conceptual framework of computations in mid-level vision

Jonas Kubilius[1,2]*, Johan Wagemans[2] and Hans P. Op de Beeck[1]

[1] Laboratory of Biological Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium
[2] Laboratory of Experimental Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

Edited by:
Antonio J. Rodriguez-Sanchez, University of Innsbruck, Austria

Reviewed by:
Jonathan W. Peirce, Notthingham University, UK
Heiko Neumann, Ulm University, Germany

***Correspondence:**
Jonas Kubilius, Laboratories of Biological and Experimental Psychology, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102 bus 3714, Leuven 3000, Belgium
e-mail: jonas.kubilius@ppw.kuleuven.be

If a picture is worth a thousand words, as an English idiom goes, what should those words—or, rather, descriptors—capture? What format of image representation would be sufficiently rich if we were to reconstruct the essence of images from their descriptors? In this paper, we set out to develop a conceptual framework that would be: (i) biologically plausible in order to provide a better mechanistic understanding of our visual system; (ii) sufficiently robust to apply in practice on realistic images; and (iii) able to tap into underlying structure of our visual world. We bring forward three key ideas. First, we argue that surface-based representations are constructed based on feature inference from the input in the intermediate processing layers of the visual system. Such representations are computed in a largely pre-semantic (prior to categorization) and pre-attentive manner using multiple cues (orientation, color, polarity, variation in orientation, and so on), and explicitly retain configural relations between features. The constructed surfaces may be partially overlapping to compensate for occlusions and are ordered in depth (figure-ground organization). Second, we propose that such intermediate representations could be formed by a hierarchical computation of similarity between features in local image patches and pooling of highly-similar units, and reestimated via recurrent loops according to the task demands. Finally, we suggest to use datasets composed of realistically rendered artificial objects and surfaces in order to better understand a model's behavior and its limitations.

Keywords: mid-level vision, similarity, pooling, perceptual organization, summary statistics

VISION AS AN IMAGE UNDERSTANDING SYSTEM

The visual system of primates processes visual inputs incredibly rapidly. Within 100 ms observers are capable of reliably reporting and remembering contents of natural scenes (e.g., Potter, 1976; Thorpe et al., 1996; Li et al., 2002; Quiroga et al., 2008). Such fast processing puts tight constraints on models of vision as most computations should be done roughly within the first feed-forward wave of information. Efforts to understand how this is possible have led to the so-called standard view of the primate visual system where objects are rapidly extracted from images by a hierarchy of linear and non-linear processing stages, where simple and specific features are combined in a non-linear fashion, resulting in increasingly more complex and more transformation-tolerant features (Fukushima, 1980; Marr, 1982; Ullman and Basri, 1991; Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; DiCarlo et al., 2012; see Kreiman, 2013, for a review).

In particular, in primate visual cortex the earliest stages of visual processing are thought to act as simple local feature detectors. For example, retinal ganglion and lateral geniculate nucleus cells preferentially respond to blobs with center-surround organization (Kuffler, 1953; Hubel and Wiesel, 1961), while neurons in primary visual area V1 respond to oriented edges and bars (Hubel and Wiesel, 1962; see Carandini et al., 2005, for a review). These detectors act locally (within their receptive field) and thus are very sensitive to changes in position or size. In contrast, neurons in the final stages of visual processing in the inferior temporal cortex respond to complex stimuli, including whole objects (Tanaka, 1996; Kourtzi and Kanwisher, 2001; Op de Beeck et al., 2001; Huth et al., 2012), faces (Desimone et al., 1984; Kanwisher et al., 1997; Tsao et al., 2006), scenes (Epstein and Kanwisher, 1998; Kornblith et al., 2013), bodies (Downing et al., 2001; Peelen and Downing, 2005) and other categories. At this stage, neurons have large receptive fields and thus are tolerant to changes in position, size, orientation, lighting, and clutter (DiCarlo and Cox, 2007). While the exact details of the properties of neurons at the low and high visual areas remain an area of active research, in our view the most puzzling question is the following: What computations are performed at the intermediate steps of information processing in order to bridge simple local early representations to highly multidimensional representations of objects and scenes?

In primates, inspired by Hubel and Wiesel's (1965) proposal of the hierarchical processing in the visual cortex, a number of studies focused on demonstrating sensitivity to the increasing complexity of features along the visual hierarchy. For example, in V2 angle or curvature detectors have been reported (Dobbins et al., 1987; Ito and Komatsu, 2004). In V4, neurons are sensitive to even more complex curved fragments and three-dimensional parts of surfaces (Pasupathy and Connor, 1999, 2001, 2002; Yamane et al.,

2008). Thus, the idea is that intermediate layers are responsible for gradually combining simpler features into more complex ones (Riesenhuber and Poggio, 1999; Rodríguez-Sánchez and Tsotsos, 2012).

However, building a system that could robustly utilize such a connection scheme on natural images is difficult. On the one hand, combining simpler features into more complex ones is complicated due to the presence of clutter. Robust mechanisms are necessary to combine the "correct" features and leave out the noise. Similarly, in order to detect complex features, enormous dictionaries must be built since the number of possible feature combinations is huge, so this process is highly resource-intensive (but see Fidler et al., 2009, for an inspiring approach to the issue). On the other hand, focusing solely on edges and their combinations into shapes misses a number of other useful cues in the images—such as differences in color, texture, motion and so on—and thus may lack the necessary power both to process object shapes and to be useful for other tasks that the visual system is performing (e.g., interaction with objects in a scene, navigation, or recovering spatial layout; Regan, 2000).

Thus, in computer vision, partially due to the described limitations of the standard view of primate visual system and partially due to the development of robust algorithms for dealing with large numbers of features, the actually implemented models of vision have bypassed thinking about intermediate representations altogether in their implementations. Instead, such models rely solely on the established features of V1 (namely, oriented edge detection) and directly apply sophisticated machine learning techniques (such as support vector machines) to detect what object categories are likely to occur in the given image. Somewhat surprisingly, this idea works very well for a number of complex tasks. For example, in the famous algorithm by Viola and Jones (2001), faces are detected using several simplistic feature detectors, reminiscent of the odd and even filters of V1. In Oliva and Torralba's GIST framework (2001, 2006; Torralba and Oliva, 2003), scene categorization is achieved by computing global histogram statistics of oriented filter outputs. Flat architectures of SIFT (Lowe, 2004) or HoG (Dalal and Triggs, 2005) that largely rely on oriented feature detection have seen a wide adoption for a variety of visual tasks in computer vision, and, in combination with multi-scale processing (Bosch et al., 2007), for a long time these models that have no hierarchies have been the state-of-the-art approach.

However, eventually hierarchical models that contain intermediate representations ultimately proved superior in many complex visual tasks. While such deep networks have been proposed several decades ago, (Fukushima, 1980; LeCun et al., 1989; Schmidhuber, 1992), only recently upon development of more robust procedures for learning from large pools of data (Hinton and Salakhutdinov, 2006; Boureau et al., 2010) such networks managed to achieve state-of-the-art object identification performance on demanding datasets that contain millions of exemplars, such as the Large Scale Visual Recognition Challenge (Deng et al., 2009; Krizhevsky et al., 2012; Sermanet et al., 2013; Szegedy et al., 2014), or that demand fine-grain discrimination as in the case of face recognition (Lu and Tang, 2014; Taigman et al., 2014). Moreover, these networks have been reported to perform extremely well on a number of visual tasks (Razavian et al., 2014). While many challenges remain (Russakovsky et al., 2013), the fact that base-level object categorization and localization have been very successful and in some cases even approaching or superseding human-level performance (Serre et al., 2007; Lu and Tang, 2014; Taigman et al., 2014) is greatly encouraging. Importantly, representations learned by such deep networks have been shown to match well the representations in the primate V4 and IT (Yamins et al., 2014), demonstrating the relevance of these models to understanding biological vision.

Naturally, the success of these object recognition models begs the question whether we now understand how the visual system processes images. It is tempting to conclude that weakly organized collections of features are sufficient for object and scene categorization, and, by extension, scene understanding. However, it is important to realize that, engineering advances aside, each layer in these architectures is based on the same principles characterized in the early visual processing of the primate brain. Is there really nothing more going on in the intermediate stages of processing?

In the following section, we consider what the computational goal of mid-level vision might be (cf. Marr, 1982). Based on these insights, in Section "Intermediate Computations" we propose basic computational mechanisms that we hypothesize to be sufficient to account for processes occurring at intermediate stages. Finally, we discuss what model evaluation procedures could help in guiding the implementation of such a system.

WHAT DO MID-LEVEL VISUAL AREAS DO?

FEATURE INTERPOLATION

Typically, a model of vision is operationalized as a feature extraction system. Features that are present in the input image need to be detected, so that a veridical (or at least useful) representation of the world (or objects in it) can be reconstructed. However, visual inputs are necessarily impoverished (e.g., due to collapsing of the third dimension as the image is projected on the retina), incomplete (e.g., due to some objects partially occluding others), ambiguous (e.g., due to shadows), and noisy. As a consequence, the problem of vision is not only feature detection but also feature inference (Purves et al., 2014).

A number of studies have shown that mid-level vision is heavily involved in feature inference. Consider, for example, the seminal series of studies by von der Heydt et al. (1984), von der Heydt and Peterhans (1989), who compared neural responses to the typical luminance-defined stimuli and the neural responses to the same stimuli defined by cues other than luminance. In one of their conditions, a stimulus was composed of two regions containing line segments but with one region shifted with respect to the other, forming an offset-defined discontinuity in the texture, which we refer to as a second-order edge (**Figure 1A**). Importantly, a simple edge-detecting V1 model would not be able to find such edges, so if some neurons in the visual cortex were responding to such stimuli, it would mean that a higher-order computation is at work that somehow is capable of integrating information across the two regions in the image.

Consistent with the known properties of early visual areas, the researchers observed a robust response to the luminance-defined edges. However, in addition they also demonstrated that

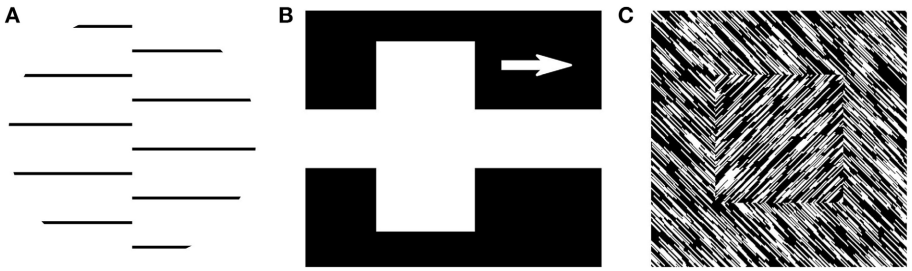

FIGURE 1 | Feature interpolation. (A) A second-order boundary stimulus as used by von der Heydt et al. (1984). **(B)** A stimulus with an illusory contour is perceived in the white gap between the two parts of the white rectangle, as used by von der Heydt et al. (1984). The arrow indicates that the white rectangle was moving. **(C)** A stimulus where a shape is defined entirely by second-order cues (that is, a difference in orientation), used in many figure-ground segmentation studies (e.g., Lamme, 1995).

some neurons in V2 responded to the second-order edges, and, in fact, often with the same orientation preference as to luminance-defined edges. Moreover, Lamme et al. (1999) reported that V1 neurons were also responding to this boundary roughly 60 ms after stimulus onset and suggested iso-orientation suppression as a mechanism behind such fast second-order edge detections. These findings have since been replicated in V2 and V4 (Ramsden et al., 2001; Song and Baker, 2007; El-Shamayleh and Movshon, 2011; Pan et al., 2012) and also reported for discontinuities in orientation (Larsson et al., 2006; Allen et al., 2009; Schmid et al., 2014), motion (Marcar et al., 2000), and contrast (Mareschal and Baker, 1998; Song and Baker, 2007; Li et al., 2014). Taken together, these findings demonstrate that even in the absence of luminance-defined borders in the inputs, mid-level areas infer potential borders from differences in other cues. Importantly, this operation is different from the typical feature detection and combination scheme because in this case a feature is computed that is not present in the input (that is, a second-order border).

An even more extreme example of such feature inference has been demonstrated by another condition in von der Heydt and colleagues' experiments where they used a stimulus inspired by the Kanizsa triangle (Kanizsa, 1955). The stimulus was defined as a white bar moving over two black bars, separated by a white gap (**Figure 1B**)—thus, although physically there were no edges connecting the two halves of the white bar, subjectively observers would nonetheless report seeing the complete white bar, effectively interpolating its borders or surface across the white gap. We refer to such borders as illusory contours. Surprisingly, for this condition, von der Heydt et al. (1984) also reported neurons in V2 responding to these illusory contours, and, in fact, nearly as vigorously as to the luminance-defined ones.

If these examples appear only as curious cases of feature inference in artificial setups, imagine a typical cluttered image where multiple objects are partially occluded. Just like in the two previous cases, the visual system appears to interpolate occluded parts of objects at the early stages of visual information processing (a process known as amodal completion; van Lier et al., 1994; Ban et al., 2013). For example, **Figure 2A** is interpreted as a gray blobby shape partially occluded by the black blobby shape, both on a dotted background, as in **Figure 2C**. In fact, we cannot help but perceive the gray shape inferred behind the black occluder and our phenomenology is most certainly not captured by segmentation into separate non-overlapping regions as in **Figure 2B**.

Similarly, the background appears to continue behind the two shapes even though there is no physical connection between the left and the right portion of it, demonstrating that filling-in is not confined to objects but applies in a more generic manner to any occluded region in the input. Moreover, at least phenomenologically, this filling-in appears to involve not only surface interpolation but also the spread of feature statistics. In our example, observers would report that the occluded part of the background is likely to continue the pattern of polka dots (van Lier, 1999).

Moreover, just like in the other two cases (second-order borders and illusory contours), the amodal interpolation has been reported to be established relatively fast, already in 75–200 ms (Sekuler and Palmer, 1992; Ringach and Shapley, 1996; Murray et al., 2001; Rauschenberger et al., 2006), and has also been observed in the early modulation of the occluded parts of shapes in monkey V4 (Bushnell et al., 2011; Kosai et al., 2014).

Taken together, we see that the visual system actively performs feature inference and it is an early process that may be initiated already with the first wave of information. It is important to note that in all of these cases, the inference does not necessarily produce a complete feature or a shape. Rather, it may reflect a rough estimate of statistical properties of the shape (cf. "fuzzy completions," van Lier, 1999) or the probability of possible completions where the missing part of the shape may occur (D'Antona et al., 2013).

RELATIONAL INFORMATION AND SURFACE CONSTRUCTION

But what is the purpose of feature extraction or interpolation? In many object recognition models, for example, the extracted features are used directly to perform categorization. Notice that such an output lacks the explicit assignment of the features to one object or another, that is, object shapes are not explicitly represented. Such model behavior is strikingly at odds with our phenomenology dominated by explicit object shapes or surfaces. This idea has been nicely illustrated by Lamme (1995) who investigated neural responses to a shape entirely defined by a second-order boundary. His stimulus consisted of a field of oriented noisy elements embedded in a background of an opposite orientation (**Figure 1C**). In order to perceive this shape, the visual system

FIGURE 2 | Seeing is not the same as perceiving. Observers report perceiving the configuration in **(A)** to be composed of full shapes as depicted in **(C)** rather than as in **(B)** which reflects the physical inputs where shapes are fragmented and two portions of background are separate. In **(C)**, the gray shape has been interpolated behind the black shape (depicted in green), indicating that mapping of a two-dimensional surface in a three-dimensional space is already necessary to represent depth relations. Furthermore, the background is also a single surface rather than two separate regions and also with its statistical properties (polka dot pattern) filled in. Some observers will also see the black shape interpolated behind the gray one (depicted in red), but this percept is much less consistent among observers than the completion of the gray shape behind the black one, indicating that surface inference might not be precise and rather indicate probabilities of possible contour and surface properties.

must be able to (i) infer second-order borders and (ii) combine them into the shape as a whole. Lamme (1995) showed that neurons in monkey V1 with receptive fields inside that shape reliably respond more than those outside, that is, the visual system explicitly represents where the figure is. Moreover, the observed enhancement was not instantaneous but rather developed in three stages (as described in Lamme et al., 1999). Early on, only responses to local features were observed. Within a 100 ms, responses to the second-order boundary emerged. Finally, neurons in V1 corresponding to the figural region of the display started responding more than the background. This effect was later shown to be the effect of feedback from higher visual areas such as V4, where such figure-ground assignments are thought to emerge (Poort et al., 2012). Taken together, this example demonstrates that the visual system gradually extracts not only the contour of a shape but also its inside, resulting in a full surface reconstruction.

More broadly, it has been argued that surface-based representations form a critical link between early- and high-level computations (Nakayama et al., 1995; see also Pylyshyn, 2001). Moreover, the presence of a surface strongly influences even the earliest computations of the visual information processing such as the iso-orientation suppression (Joo and Murray, 2014). Finally, surface-based representations can also be beneficial for object identification tasks because surfaces are topologically stable structures and thus largely invariant to affine transformations (Chen, 1982, 2005). For example, a hole in a surface remains present despite drastic changes in its position, orientation or rotation in depth, or to the changes in surface structure (Chen, 1982; Todd et al., 2014).

In general, we argue that encoding spatial relations—whether between features, or deciding which features belong to the same object or surface, or ordering the surfaces in space—provides a tremendous wealth of information (Biederman, 1987; Barenholtz and Tarr, 2007; Oliva and Torralba, 2007): Knowing that a car is on the road or above the road makes a big difference, but using only features without relations between them might fail to capture these differences (Choi et al., 2012). One influencial account of the power of spatial relations has been provided by Biederman (1987), who noticed that certain spatial relations between features, known as non-accidental properties, remain largely invariant to affine transformations in space. For example, short parallel lines nearly always remain parallel despite changes in viewpoint. He proposed that these relations might be used to encode different object categories, and later Hummel and Biederman (1992) developed a model illustrating how such a system might work. While the exact purpose of such structural representations in recognition has been heavily debated since (Barenholtz and Tarr, 2007), consistent with this idea a number of studies demonstrated that observers are very sensitive to changes in these invariant features of a shape (Wagemans et al., 1997, 2000; Vogels et al., 2001; Kayaert et al., 2005a,b; Lescroart et al., 2010; Amir et al., 2012).

Similarly, Feldman (1997, 2003) and van Lier et al. (1994) argued that configural regularities of the inputs are used to organize features into objects, and human visual system has been shown to be sensitive to such configural relations (Kubilius et al., 2014). Moreover, Blum (1973) proposed that the configuration of shapes is encoded in the visual system by representing their skeletal, or medial axis, structure, and Hung et al. (2012) showed that neurons in monkey IT indeed respond both to the contour of a shape and its medial axis structure. Taken together, these studies highlight the fact that the visual system utilizes configural relations between features and surfaces in the higher visual areas, and therefore an explicit encoding of these relations should be supported by mid-level computations.

REPRESENTATIONS FOR MULTIPLE TASKS, NOT ONLY OBJECT RECOGNITION

We argued that mid-level vision was involved in feature detection and surface construction, such that in the end the shape of an object could be reliably extracted from the image. However, the long quest for superior object identification algorithms has somehow overshadowed the fact that visual cortex can achieve more than just object identification. Vision is our means to understanding the world, whereas a mere object-based representation provides only a tiny fraction of information needed for successful behavior in the world. This point is particularly pertinent in lower species such as rodents for whom navigation is a more immediate task than object identification (Cox, 2014). In fact, much

of our visual input is not composed of well-defined objects and thus trying to parse them into objects makes little sense. A richer description is thus needed if we were to capture the essence of information about the world (Gibson, 1979).

To stress the point of the inadequacy of object-based representations, let us consider a series of images in **Figure 3**. In some cases, like **Figure 3A**, where the object ("a car") is clearly separate (self-contained) from the rest (the road), object identification and localization provides the most important information about the scene ("there is a car"). But consider a row of buildings, for example (**Figure 3B**). While one still clearly describes each house as a distinct object, they are impossible to detach from other items (other houses and the ground). A more extreme example is depicted in **Figure 3C**, where even though a mountain is sticking out from the ground surface, it is no longer very clear where the mountain ends and the ground begins. Is the visual system really concerned about finding objects in such images then? In fact, as we go further away from close-up views into panoramic scenes, identifying objects does not appear to be the default mode any longer. In **Figure 3D**, we know that the image is composed of individual trees, grass and other stuff but we no longer can count them. Rather, a percept of various textures and layouts appears to dominate. Thus, talking about individual objects is largely irrelevant in these scenarios and instead describing texture properties and characteristics that allow navigation through the terrain, or a global level semantic labeling of "a forest" or "a lawn" often seems to be the more immediate task for vision (Oliva and Torralba, 2001; Torralba and Oliva, 2003).

Therefore, we point out that surfaces that mid-level areas construct are not only meant to represent the outline of objects in images but also (or primarily) to summarize the properties of textures and surfaces in the environment.

REPRESENTATIONS PRIOR TO IDENTIFICATION

Finally, we point out that intermediate representations do not have to rely on being able to identify the contents, consistent with the idea that they are computed early on. We do not need to know what we are looking at to be able to describe its three-dimensional shape, texture, and spatial relations to other items in an image. For example, notice that in **Figure 2** surface interpolation occurs despite us never having seen these particular shapes before and having no categorical label for them, indicating that this phenomenon could be performed by mid-level computations prior to categorization. This observation also holds for a more realistic image depicted in **Figure 4**, where we can easily agree that five objects situated in different depth planes are depicted. We can describe their shape and imagine acting upon them despite partial occlusions present in the image. This is clearly a more advanced representation of the image contents than a mere V1 filter output, yet not so advanced as to require any categorization, recognition or identification (naming) of the objects in it.

The idea of intermediate representations being established without recognition of contents is well-known in psychology (Witkin and Tenenbaum, 1983; Nakayama et al., 1995). To provide an illustrative example, the famous visual agnosia patient DF cannot report the identity or even orientation of most objects, yet her ability to act on these objects remains intact, a finding that has led Goodale and Milner (1992) to propose the vision-for-action and vision-for-perception division in the visual information processing in the brain. It thus appears that our visual system is adept in processing inputs even lacking knowledge about what they are, pointing to the idea that scene segmentation into objects might be more basic or more immediately performed than recognition. We do not claim that recognition is irrelevant for segmentation, as it has been shown that recognition can bias figure-ground assignment (Peterson, 1994), but our point is that it can largely be done successfully without any knowledge about the identity of objects.

CONCLUSION

Taken together, we claim that the goal of mid-level areas is the construction of surface-based representations that segment the input images into objects, background surfaces, and so on, together with their textural properties, because such format of representations is sufficiently rich for the variety of high-level tasks, including three-dimensional reconstruction of the scene, navigation in it, interaction with objects or restricting attention to them. The idea of the primacy of the surface-based representation is also supported by empirical studies showing that some form of figure-ground organization would be established already shortly after feedforward inputs reach higher visual areas and is consistent with the observation that segmentation does not require knowledge of the identity of the objects involved. Importantly, given the computational complexity, this organization is probably not computed globally but rather is restricted to parts of visual inputs that fall at fixation or where an observer is attending.

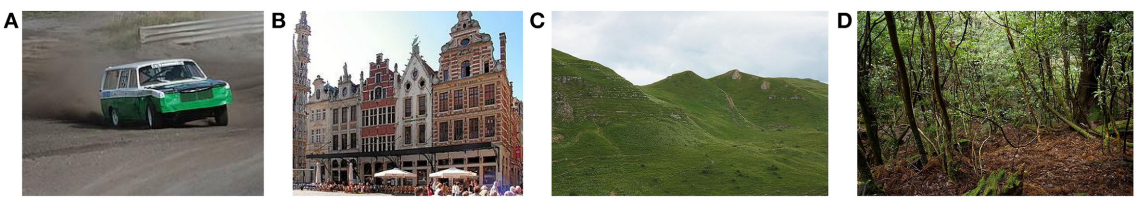

FIGURE 3 | The hierarchy of objecthood. Objects are not the most important piece of information in every image. While **(A)** has a well-defined object, it is already less clear in **(B)** what should count as one: The row of houses? Or each house separately? Or each of the windows? In **(C)**, there are three mountains but where each of them begins and ends is neither clear nor very important, and in **(D)** layout rather than object identity dominates perception, although one can see trees, trunks, etc. (Image credits from left to right: bengt-re, 2009, Snowdog, 2005, Reza, 2009, Σ64, 2012. All images are available under the Creative Commons Attribution License or are in the public domain.).

FIGURE 4 | Recognition is not crucial for scene or object understanding. In this artificially generated scene we see five novel objects, we can describe their three-dimensional shape despite partial occlusions, and navigate around them without having to know the identity of those objects.

It is also important to understand that the segmentation we describe here is not the same as what is commonly meant by this term. Many algorithms of segmentation only divide the image into a mosaic of non-overlapping regions without any information about the depth, that is, which region is in front of another one (see also Section "Current Approaches"). However, whenever something is occluded, that is a cue for depth ordering. Therefore, we consider a process that not only divides the image into separate regions but also infers figure-ground relations between these regions. Since this process often involves the inference of occluded parts, we refer to such interpolated regions as a surfaces.

Finally, such depth ordering is necessarily an oversimplification. For example, observe in **Figure 2C** that we do not perceive the whole of the black shape in front of the gray one. In fact, at least for some observers, part of the black shape (shown in red in **Figure 2C**) appears to be behind the gray shape, suggesting a three-dimensional form of the two shapes (Tse, 1999). This example demonstrates that the resulting representations cannot be captured by splitting an image into several depth planes, and thus require more flexibility. Such representation presumably would be followed by a full rectification of a three-dimensional volume at the later stages of visual information processing.

INTERMEDIATE COMPUTATIONS

We proposed that intermediate processing stages produce surface-based representations from two-dimensional static images. What computations could produce such representations?

CURRENT APPROACHES

In computer vision, many early image segmentation approaches considered segmentation as a global optimization problem of finding the best boundaries, grouped regions, or both. For example, Mumford and Shah (1989) proposed a functional that estimates the difference between the original image and its segmentation with constraints for smoothness and discontinuity at region boundaries (see also Lee et al., 1992). Finding the best segmentation amounts to finding the global minimum of this functional. Similarly, in a boundary-based contour extraction model, Elder and Zucker (1996) considered finding the shortest-path cycles in the graph containing boundary elements.

However, solving for a global optimum deemed to be a complicated task, often leading to unsatisfactory results. In 2000, Shi and Malik proposed a reconceptualization of the image segmentation problem as a graph cut problem. When features in an image are represented in a graph, finding the best segmentation amounts to finding groups of features in this graph that are maximally similar within a group and maximally dissimilar from other groups. Shi and Malik (2000) showed that their normalized cuts algorithm could provide a good optimization of this criterion and, based on this approach, they later developed one of the best-known image segmentation models (Arbeláez et al., 2011; see also Felzenszwalb and Huttenlocher, 2004 and Sharon et al., 2006, for much faster implementations of this idea).

Partitioning a graph in a fixed way, however, cannot capture the inherently hierarchical structure of images (a part can be part of another part; see the windows of houses in **Figure 3B**), nor can it adapt to the task demands. Therefore, in recent years much effort in image segmentation research has been devoted to the development of methods for the probabilistic generation of region proposals (Arbeláez et al., 2014) that could then be refined using a higher-level task such as categorization (Leibe et al., 2008; Girshick et al., 2014; Hariharan et al., 2014) or would be flexibly reconfigured based on Gestalt principles (Ion et al., 2013).

How could such partitioning of an image graph into high-similarity clusters be implemented in a biologically-plausible architecture? Based on behavioral and neural evidence, Nothdurft (1994) hypothesized that image segmentation involves (i)

suppression of responses in homogenous feature fields, and (ii) local pooling of features for boundary detection. Unlike the global optimization approaches considered above, this idea is based on completely local computations that are attractive due to their low complexity and biological plausibility. The implementation of this idea can be found in models by Grossberg (1994) and Thielscher and Neumann (2003), where texture segmentation is performed by enhancing edges that group together by the good continuation cue (using the "bipole cell" idea), and suppressing other locations in the image. Repeated over several iterations, this computation leads to the formation of the outline of the shape. This idea accounts well for Nothdurft's (1994) observations, and also provides an integrated framework of using both texture and boundary information to perform segmentation. Moreover, Thielscher and Neumann (2005) also demonstrated that this approach produces differences in convex and concave boundary appearance, in line with Nothdurft's (1994) observations.

Segmentation into distinct regions is only the first step though. As discussed in the previous section, this is not sufficient because an explicit surface construction and figure-ground relation computation need to occur as well. Some approaches (Roelfsema et al., 2002) attempted to explain figure-ground segmentation simply as an effect of increasing receptive field sizes (thus, decreasing spatial resolution) in higher visual areas. The model operates by initially detecting boundaries in the inputs and then pooling them together in higher visual areas as a result of increasing receptive-field sizes. Eventually, the whole shape is represented by a unit with a sufficiently large receptive field. Then, the figure-ground assignment can be propagated down via feedback to the early visual areas, as observed in the experiments by Lamme (1995).

However, it is unlikely that such scheme would work in more complex displays with more overlapping shapes and more variation in texture. Moreover, smaller shapes always produce higher responses in higher-level areas because their boundaries are closer together. Since these responses represent the figure-ground signal, smaller shapes are always bound to be on top of larger shapes that produce a weaker figure-ground signal. One possibility to resolve some of these issues is to use corners as indicators of the figural side. Since figures tend to be convex, the inside of a corner reliably indicates the boundary of a figure. Based on this observation, Jehee et al. (2007) proposed an extended version of the model by Roelfsema et al. (2002) that could produce more reliable border assignments.

The idea of using convexity can be applied more generally across the entire shape outline and not only at its corners. To illustrate how that could work, consider the two shapes in **Figure 5A**. The two edges shown in red can either be the boundary of the gray surface or the boundary of the white one, as indicated by the green arrows pointing to both directions. Of course, in this case it is clear that these edges must belong to the gray surface because the white one is just the background. But how would a model know? If we assume that objects tend to be convex, edges that are in agreement (the green arrows that are pointing toward each other) might belong to the same surface (**Figure 5B**). This simple computation in the local neighborhood followed by pooling into curved segments (**Figures 5C,D**) results in a largely correct border ownership. If it is further computed globally over a few iterations, local inconsistencies (e.g., a concavity of the lighter gray object) can be resolved (**Figure 5E**; see Figure 5B in Craft et al., 2007, for a working example), resulting in the proper assignment of edges to one of the two objects (**Figure 5F**), which is the desired initial image division into surfaces.

Importantly, because of border-ownership, we also learn which parts of objects are occluded. If a certain surface is partially bounded by a boundary that it does not own, it is a sign of an occlusion. For example, in **Figure 5F**, the yellow object is partially occluding the blue one, and border-ownership assignment indicates that edges along the yellow object belong to it. That leaves the blue object lacking a closed contour, meaning that part of it is occluded. An interpolation of surface results in a more perceptually compelling segmentation into whole shapes (van Lier et al., 1994), and consequently provides an ordering of surfaces in depth (**Figure 5G**).

The existence of such border-ownership cells has been reported in the visual area V2 (Zhou et al., 2000; see Zucker, 2014, for a good overview) and a number of models based on this idea have been proposed since (Zhaoping, 2005; Craft et al., 2007; Layton et al., 2012). Kogo et al. (2010) extended this framework by also using L- and T-junctions to determine not only figure-ground assignment for luminance-defined figures but also to produce the correct output in the case of illusory contours (Kanizsa's figures). Importantly, unlike earlier proposals (e.g., Grossberg, 1994), their approach is capable of yielding the correct representations of comparable yet non-illusory displays without ad hoc deletion of interpolated contours (see **Figure 1B** in Kogo et al., 2010).

Similarly, extending their work on bipole cells, Thielscher and Neumann (2008) showed that T-junctions could be used to infer figure-ground relations for multiple figures (not just figure and ground) in their architecture, and more recently, Tschechne and Neumann (2014) extended their earlier work to a full model of figure-ground segmentation. Initially, bipole cells, curvature and corner detectors are used to produce the consistent outline of a shape. Then, contextual cues are used to compute border-ownership.

Taken together, current biologically-inspired approaches to image segmentation largely concentrate on discovering boundaries in an input image and resolving figure-ground assignment by computing border-ownership of the boundaries in an image. However, unlike purely computer vision algorithms, these approaches are typically not tested with realistic inputs, thus their applicability and robustness on the wide variety of natural images remains unclear. Moreover, some models are better at segmentation but do not perform feature interpolation and figure-ground relation computations, and vice versa, while others focus on using second-order features but are not robust for segmentation using multiple cues, and so on. In other words, each of them only implements several aspects of processes in mid-level vision but the proposed mechanisms are not mutually compatible to build a unified architecture. Could there be several basic mechanisms that could account for the majority of the available data?

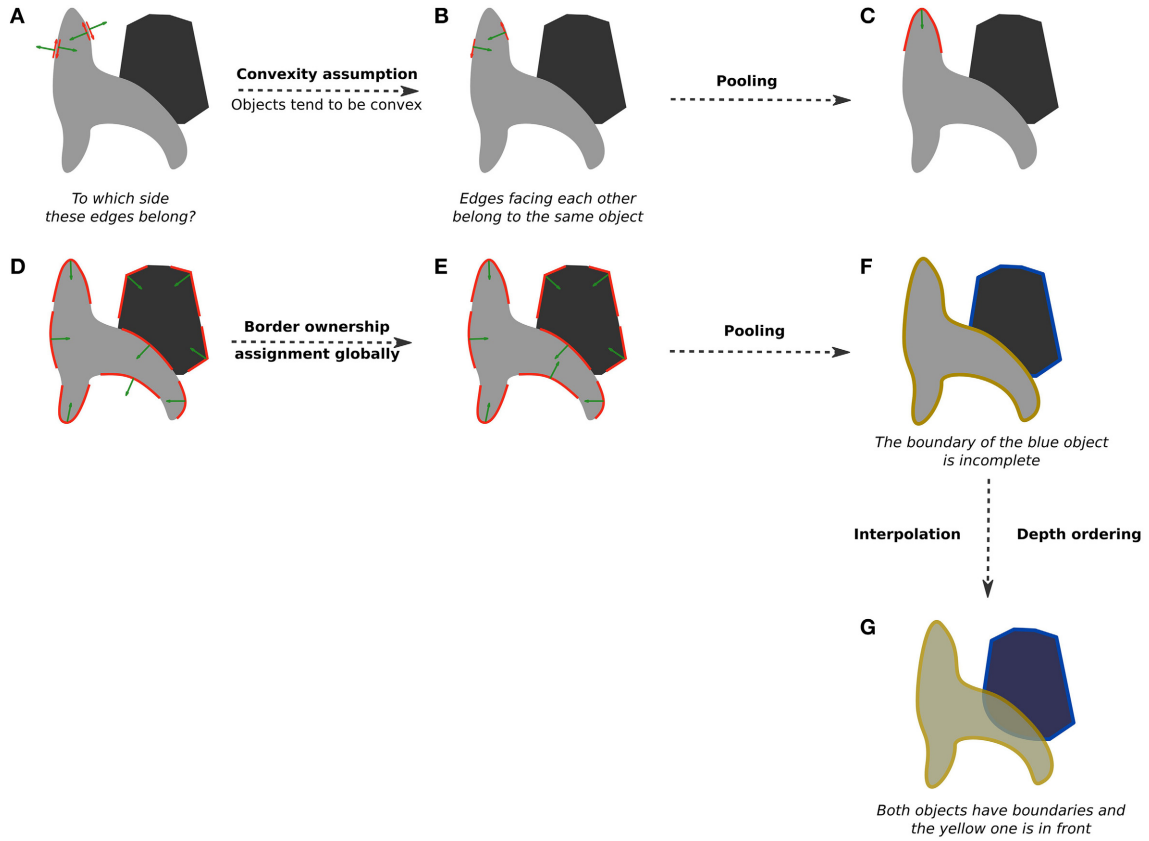

FIGURE 5 | An illustration of border-ownership assignment. (A) Initially, each edge in the image (red arrows) can belong to one of two sides (green arrows): either the gray surface or the white surface. The goal of border ownership computation is to figure out which side is the figural side. In this case, the edges should belong to the gray surface. **(B)** Using the convexity assumption (objects tend to be convex), we can easily determine border ownership in the local neighborhood. Edges that "agree" (green arrows are "looking" at each other) are preferred. **(C)** Pooling these edges together results in a curved segment with the correct border-ownership assignment. **(D)** After this computation is carried out in the local neighborhood, border ownership is largely but not fully correct. We can improve it by using the same convexity assumption over larger areas (e.g., over the entire image). **(E)** Global border-ownership computation results in a correct assignment of all segments. **(F)** With pooling, two separate surfaces emerge. Note that the blue one is missing a boundary at the intersection with the yellow object. This implies that the blue object is partially occluded by the yellow one. **(G)** Using this information, a correct local depth ordering is established and the missing piece of the blue object is interpolated.

OUR APPROACH

In a nutshell, we are interested in understanding *conceptually* what computations could suffice to account for the following biologically-plausible image processing strategy:

1. Region property and boundary extraction.
2. Clustering of boundary and region features into separate surfaces (segmentation).
3. Surface interpolation and depth ordering (figure-ground organization).
4. Representation refinement via recurrent loops.

Moreover, we want these computations to be sufficiently robust such that they would apply across various features in the images and could therefore be used in the typical computer vision setups such as deep networks.

To implement steps 1 and 2, we propose two basic mechanisms for intermediate computations, generalizing the vast majority of approaches discussed in Section "Current Approaches" (**Figure 6**):

- *similarity statistics* that compute correlations between local patches of the input, and
- *pooling* that combines together highly similar (well-correlated) patches.

These two computations are implemented hierarchically, processing over increasingly larger patches of the input image and resulting in a coarse mid-level representation of surfaces and their properties upon the first roughly feedforward processing wave. As a result of feature inference at multiple layers, the constructed surfaces are partially overlapping, providing information for depth ordering at the highest stages of this architecture (step 3). The resulting representations will be very coarse and probably inconsistent, so an iterative refinement of these representations by reapplying similarity and pooling operations over

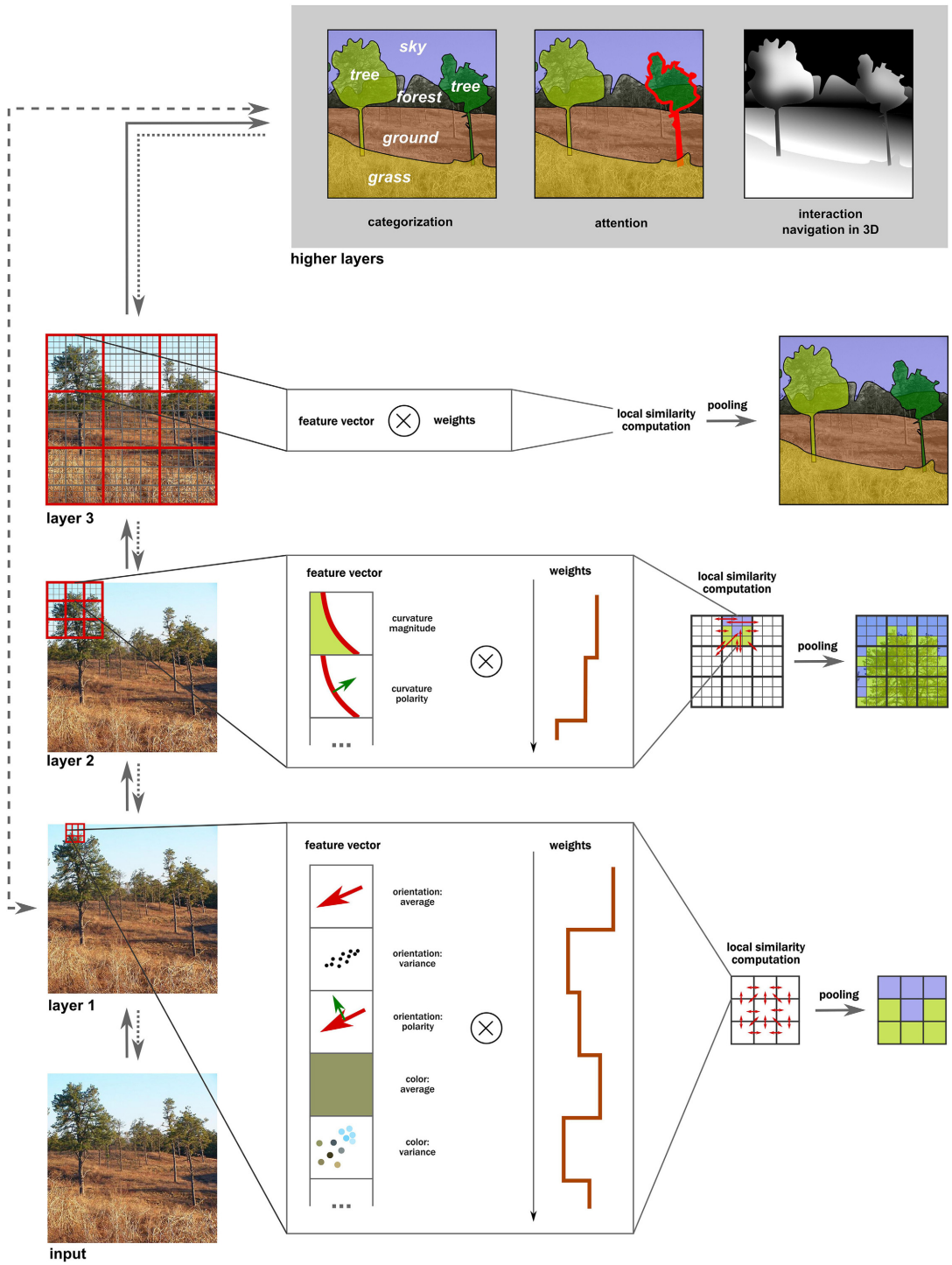

FIGURE 6 | Computation of intermediate representations in the visual hierarchy. In each layer, various features are extracted first at each location, forming a feature vector. Next, correlations are computed in the local neighborhood between each pair of a weighted feature pair, leading to similarity statistics (red arrows). (The optimal weights need to be learned by training the model.) Finally, these patches are pooled together into clusters that contain similar statistics. These new clusters are used in the next layers for the same similarity and pooling over increasingly larger neighborhoods. Note how the resulting intermediate representations are interpolated behind occlusions and are ordered in depth (e.g., the tree is in front of the forest). These representations can now be used for higher-level tasks such as categorization, attention to specific objects or interaction with them, or for navigation. They are also rather coarse initially (e.g., trees on the right are incorrectly lumped together), and can further be refined iteratively via feedback loops (if attention is directed to that region). Moreover, notice that not all steps must necessarily be carried out as certain shortcut routes (e.g., the gist computation) using simpler statistics can occur.

smaller parts of an input image is important as well (step 4; see also Wagemans et al., 2012b). We briefly discuss the role of feedback in Section "The Dynamic Nature of Intermediate Representations."

SIMILARITY ESTIMATION AND POOLING

Let us start by considering the output of a typical low-level computation such as edge detection, as illustrated in **Figure 5A**. The red arrows in this figure show the locations and orientations of salient edges in the image. While this is a useful description of potential boundary positions in the image, this information does not suffice to understand the organization of the image contents. In particular, it does not indicate which edges are likely to define the same surface, as shown in **Figure 5B**. At this stage the system only knows about separate salient edge positions, and further processing is needed to group both boundary and textural elements into coherent surfaces.

Finding which edges might group together can be achieved with a simple *similarity* measure, such as a correlation between two locations in an image. If the similarity is high, the two edges might belong to the same smooth contour (since edges at nearby locations of a smooth curve have similar orientation) or the same surface composed of similarly oriented elements (e.g., the wood texture in **Figure 4**). In contrast, a low similarity indicates a potential discontinuity in an image, or a second-order edge, just like the one between the ground and the object in **Figure 1C**.

Of course, similarity computation need not be restricted to oriented edges only and can be applied across other properties (e.g., spatial frequency, phase bands, color) and even across summary statistics within a local patch (e.g., mean and variance of orientation). Notice that by incorporating multiple cues, this single computation of similarity among the adjacent locations provides a natural approach to dealing with both boundary and textural cues in images. In particular, wherever there is sufficient dissimilarity, textural properties are actively used to generate boundary elements that are further used to construct full surface boundaries.

Freeman et al. (2013) provided evidence that such similarity measures are indeed computed early in the visual system. They constructed synthetic textures with specific higher-order statistical dependencies, such as marginal statistics, local cross-position, orientation, scale and adjacent-phase correlations, and demonstrated that such neurons in primate V2 (but not V1) were particularly sensitive to these built-in statistics, suggesting that V2 computes similarity between features. When used in textures, such summary statistics apparently are sufficient for the synthetic generation of similar-looking textures (Portilla and Simoncelli, 2000). When used on natural images, these statistics appear compatible with percept in peripheral vision (Freeman and Simoncelli, 2011; Freeman et al., 2013) and can also account for certain effects in crowding (Balas et al., 2009) and visual search (Rosenholtz et al., 2012).

Similarity statistics alone are not sufficient, however. While they are clearly useful in providing rich descriptions of the inputs, the number of parameters in the system increases dramatically since these statistics are computed pairwise between many small patches. Maintaining all these parameters does not appear to match our phenomenology where integrated shapes or regions dominate over local fragmented interpretations. Moreover, natural scenes contain substantial redundancy and the visual system appears to take advantage of it via efficient coding strategies (Attneave, 1954; Barlow, 1961; Simoncelli and Olshausen, 2001; Olshausen and Field, 2004; DiCarlo and Cox, 2007). For instance, Vinje and Gallant (2000) demonstrated that V1 neurons use a sparse encoding scheme that matches the sparse structure of natural scenes. Other researchers have demonstrated that sparsity constraint leads to the development of simple and complex cells in computational models (Olshausen and Field, 1996; Hyvärinen and Hoyer, 2000, 2001).

It thus appears that a higher-order statistic, one that would summarize similarity statistics, is necessary. We call this computation *pooling* to reflect the idea that separate units are now pooled together according to the strength of the previously computed pairwise correlations. Computationally, such pooling operation is very simple, for example, a single-link agglomerative clustering of patches that correlate above a certain threshold (Coates et al., 2012) or mean-shift (Paris and Durand, 2007; Rosenholtz et al., 2009). The threshold can be flexible (i.e., a free parameter in the model) reflecting individual differences between participants.

While either similarity or pooling have been utilized in various formats separately by many models, exploring the power of their combination is rare. Geisler and Super (2000) showed that a similar similarity and pooling scheme could account for a number of typical perceptual grouping displays. One successful demonstration of this combination on real images was reported by Yu et al. (2014) who found that a super-pixel segmentation followed by mean-shift clustering accounted surprisingly well for visual clutter perception. In a notable example that such scheme can be both powerful and efficient even for practical applications (due to parallelization), Coates et al. (2012), using K-means and agglomerative clustering, achieved robust unsupervised learning of face features using tens of millions of natural images.

HIERARCHICAL SIMILARITY ESTIMATION AND POOLING

While it would be possible to perform similarity and pooling globally across the whole image, such strategy would be very inefficient and probably not very accurate. Instead, we propose that these computations are performed hierarchically, such that first similarity and pooling are done locally, then over somewhat larger neighborhood using the newly inferred features, and finally globally using few but rather complex features that result from these computations at earlier stages.

The initial computation of a similarity and pooling would yield longer straight or curved segments (**Figure 7A**, right). A low correlation, on the other hand, would indicate the presence of second-order edges that are formed between adjacent surfaces with differently oriented elements. For example, in **Figure 7B**, left, there is no clear edge separating the object from the ground since their overall luminance is quite similar, and thus segmentation could not be done with a simple V1-like edge detection model. The desired segmentation becomes trivial when the difference in orientation content is observed. The dominant orientation of the object is different from that of the ground and

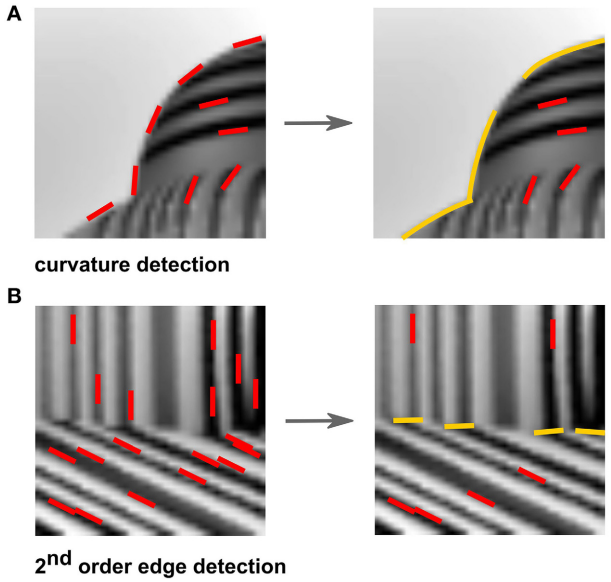

FIGURE 7 | Examples of feature inference. (A) Pooling similarly-oriented features (red) result in the inference of curved fragments (yellow). **(B)** In contrast, a dissimilarity between oriented features (red) result in the inference of second-order edges (yellow) between textures.

can therefore be used to determine a boundary between the two textures, which is indicated by the low similarity measure (**Figure 7B**, right).

Of course, detecting second-order edges in this fashion can also yield spurious results. Boundary element orientation can change significantly at inflection points (i.e., junctions) leading to low similarity measures, and yet these do not imply the presence of a second-order edge. One solution to the problem could be to use only sharp edges for defining boundaries, and otherwise assume that edges define textures (the insides of a surface). Consistent with this idea, Vilankar et al. (2014) reported that edges defining an occlusion tend to have steeper changes in contrast than non-occlusion edges (reflectance difference, surface change, cast shadows) and that a maximum likelihood classifier could predict the type of edge with 83% accuracy in their database. Another possibility is that junctions are not detected during the initial processing and only computed later when the global estimate of the shape is already available from the higher-level areas. Consistent with this idea, McDermott (2004) reported that participants were unable to report T-junctions using local natural image information (small patches of image) only (but see Hansen and Neumann, 2004; Weidenbacher and Neumann, 2009).

However, in general, the visual system is not so much interested in the features as such but in the surfaces they define. Other cues than boundaries can therefore be important in the local computations of which features should be combined into a single surface. As discussed above, convexity is an important cue for border-ownership assignment. Measuring consistency in edge polarity (where the brighter side is) can also provide information if they are likely to belong together (Kogo and Froyen, 2014). In fact, Geisler and Perry (2009) observed that edges with an inconsistent polarity are less likely to belong to the same contour. Recently, it has been reported that even low-level cues, such as the sharpness of an edge or local anisotropies in spectral power can be informative about figure-ground organization (Ramenahalli et al., 2014; Vilankar et al., 2014).

So, at each location where a boundary element has been found or inferred, we can list all these cues as a long vector and then compute the similarity between these vectors in the local neighborhood. Sufficiently similar locations are then pooled together, resulting in new, more complex features at a higher layer of this hierarchy. Now again, the similarity of these new features over larger scales can be computed, and similar features pooled together into even more complex features, such as parts of boundary (Brincat and Connor, 2006) or surface patches (Yamane et al., 2008) with a complex geometry. Finally, these features are pooled again over the entire image, producing the initial segmentation of an image into proto-surfaces.

NEURAL REPRESENTATION OF POOLED UNITS

By definition, a pooling operation combines outputs of several units and treats them as belonging to the same group (same contour, shape, or surface). Several alternatives have been proposed how such groups could be represented in the visual system. Perhaps the most straightforward way to implement this representation is by having dedicated grouping cells. Such idea has been used in a computational model of border-ownership assignment by Craft et al. (2007). They implemented neurons with donut-shaped receptive fields that can pool together units lying on that donut. However, such grouping cells have yet to be found in the visual system. It is possible however that cells with large curved receptive fields that exist in V4 might suffice to perform the border-ownership computation (as the authors themselves suggest on p. 4320 of their paper).

Another simple strategy is an increase of the mean neural response of units belonging to the same group (Roelfsema et al., 2004). However, this strategy also implies that only a single group can be maintained at a time. If another group needs to be processed, such as when shifting attention from one object to another, the integration computation would have to be performed again. While it may appear somewhat limiting, it should also be noted that in many tasks, such as multiple object tracking, observers show a rather poor ability to maintain representations of multiple groups at the same time.

A very different idea has been proposed by von der Malsburg (1981). He hypothesized that representations are held together by synchrony in neuronal firing. Such idea, if true, would in theory allow for multiple stable representations to co-exist in the visual system. While such synchrony has been observed in the visual cortex (Singer and Gray, 1995), its functional role is heavily debated, questioning whether it indeed plays a causal role in representing groups (Roskies, 1999; Roelfsema et al., 2004).

Finally, a similar idea has been put forward by Wehr and Laurent (1996). They provided evidence that locust's olfactory neurons fire in a certain unique temporal patterns to various combinations of scents. For example, while an overall response to an apple and to a mint and an apple scents might appear

comparable, at a finer temporal scale differences emerge in the number and timing of these higher frequency peaks (three peaks for the apple scent but only two for mint and apple). In other words, each stimulus receives a unique code of neural firing which can serve as a tag for belonging to a certain group. Importantly, just like binary code in computers, this code can accommodate a large number of stimuli without running into the combinatorial explosion.

THE DYNAMIC NATURE OF INTERMEDIATE REPRESENTATIONS

The visual processing need not stop with the feedforward formation of the intermediate representations. Probably the best we can expect at this first pass of processing is a very coarse representation capturing the most salient aspects of the input. For example, the initial representations may lack global consistency: it is likely that not all parts of an object will be bound into a single entity, and there can also be errors of the bounding of parts. For instance, the legs, body, and arms of a human body might be separate initially if there is not enough similarity between them. As a result, these parts may also have conflicting figure-ground assignments, such that the body is computed to be behind a chair but the legs are in front. If necessary for the task, a reconfiguration of these components could be formed iteratively until a global minimum is found, resulting in a stable percept of the configuration. For instance, the border-ownership model by Zhaoping (2005) resolves the direction of border-ownership by iteratively computing which side is more likely to be the figural side. The iterations are necessary because, for example, borders in concave parts of a shape might initially have the wrong border-ownership (toward the convex side) but over several iterations the assignment is gradually reversed since other parts of the global shape influence the decision that the concavity should be part of the whole shape. There are also cases where several interpretations are similarly plausible (e.g., the Necker cube or the vase-face figure; see Wagemans et al., 2012a), and thus iterative computations will lead to continuous switches between these interpretations.

In many cases, the refinement of representations will also be necessary. In particular, the initial representation formed in mid-level areas might only capture the gist of the input. Details will be necessarily lost due to agglomerative pooling operations. In order to extract finer details, representations in earlier layers can be reaccessed via feedback loops (indicated by backward arrows in **Figure 6**), as conceptualized by the Reverse Hierarchy Theory (Hochstein and Ahissar, 2002). Such feedback connections are abundant in the primate visual cortex and have been implicated to be important for various purposes (Felleman and Van Essen, 1991; Angelucci et al., 2002; Roelfsema et al., 2010; Arall et al., 2012). For example, intermediate representations could be used as saliency maps to direct attention to a particular part of an image or a particular feature (Walther and Koch, 2006; Russell et al., 2014). Then irrelevant inputs would be inhibited while the relevant ones would receive an enhanced weight (Mihalas et al., 2011; Arall et al., 2012; Wyatte et al., 2012), and the whole similarity and pooling computation would be repeated again. Such approach could be particularly important for resolving complicated parts of images that require high spatial resolution (Bullier, 2001), serial (or incremental) grouping of image features (Roelfsema, 2006), and could play a major role in learning features from input statistics (Roelfsema et al., 2010).

Iterative computations also provide the necessary flexibility for dealing with the inherently hierarchical composition of scenes. Consider, for example, **Figure 3B**, where all buildings could be represented by a single surface, or could be further divided into separate surfaces for each building, or even further for each window or any other detail in the image. Task demands, the mental state of an observer, and other factors can have a strong influence to the percept at any given moment. Utilizing the recurrent connections, the dynamics of the percept could be modeled in our framework by updating the pooling threshold (Sharon et al., 2006; Ion et al., 2013).

Of course, the proposed system need not be strictly hierarchical. For certain computations, it makes sense to have fast bypass routes (indicated by the dashed arrow at the top of **Figure 6**) whenever construction of intermediate representations is too slow or unnecessary, as could be the case for face detection where Viola and Jones' (2001) approach proves sufficient, or for a rapid scene categorization using the gist computation (Torralba and Oliva, 2003). Moreover, including such bypass routes naturally provides the visual system with the flexibility to both build detailed representations gradually and also to produce global impressions of the input statistics rapidly (Bar, 2004). The gist of the scene can provide informative priors (category, context, memory associations, and so on) that could guide processing and segmentation at intermediate layers (Peterson, 1994; Rao and Ballard, 1999; Oliva and Torralba, 2007).

Finally, we want to stress that although recurrent processing can improve surface representations and help in task performance, figure-ground segmentation does not require it. For example, Supèr and Lamme (2007) observed that removing most of feedback connections from higher visual areas to V1 reduced but did not abolish figure-ground perception. In fact, Qiu et al. (2007) reported that border-ownership signals emerge pre-attentively, and a purely feedforward model of figure-ground segmentation has been proposed by Supèr et al. (2010), consistent with a limited role of feedback in figure-ground assignment process (also see Arall et al., 2012, and Kogo and van Ee, 2014, for a discussion).

EVALUATING PERFORMANCE

The proposed architecture is meant to simulate the representations residing in mid-level vision. Given that this is not the final stage of the visual processing, evaluating the model's performance is not trivial. Often, models of vision are evaluated using standard object identification or scene segmentation datasets such as the ImageNet (Deng et al., 2009) or the Berkeley Segmentation Dataset (BSDS500; Arbeláez et al., 2011), where the goal for a model is to produce labels or segmentations as close as possible to the correct answers defined in that dataset. So, one simple solution for testing our architecture could be to extend it to perform one of these tasks. In this section, however, we discuss how blindly applying standard benchmarks can be misleading and highlight the need for good, carefully constructed tests and datasets that would help to detect shortcomings in the model and guide its development (Pinto et al., 2008).

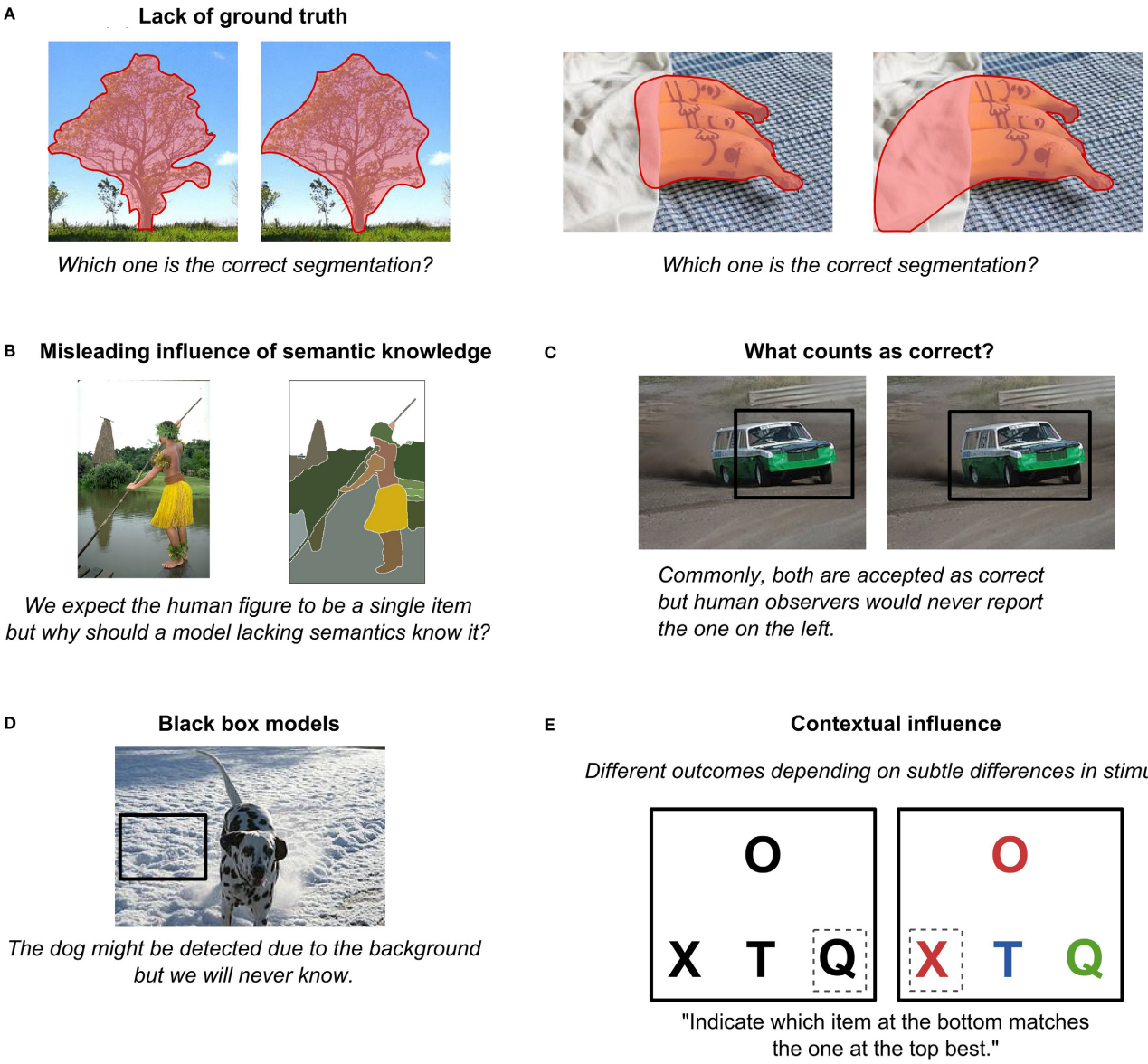

FIGURE 8 | Several potential problems with image datasets. (A) Lack of ground truth (photo credits: Sheila in Moonducks, 2010, and Camera Eye Photography, 2013). **(B)** Misleading influence of semantic knowledge (image from the dataset described in Arbeláez et al., 2011). **(C)** What counts as correct? (photo credit: bengt-re, 2009). **(D)** Black box models (photo credit: Berbezier, 2008; inspired by Landecker et al., 2013). **(E)** Contextual influence. All photos are available under the Creative Commons Attribution License.

First of all, there is always the question of the "ground truth." For example, which of the two segmentations in **Figure 8A**, left, is the ground truth? Both seem reasonable to a human observer and, in fact, they have been annotated by hand, making them, by definition, not objective. For example, smaller objects might be missing, subordinate categories might remain not annotated, and there may even be a disagreement among raters as to what constitutes an object and what is only a part of an object. While it is possible to step away from human raters altogether by obtaining ground-truth data using motion and depth information (Scharstein and Szeliski, 2002), only obtaining more precise measurements is not solving the major issue. In particular, the differences in ratings are largely driven not by imprecise annotation of boundaries but rather reflect individual differences in how people perceive images and what task they think they need to do. In other words, there is no ground truth to natural images because, as we have repeatedly pointed out in this paper, perception (and thus the definition of objects) is observer- and task-dependent. Another pertinent example to illustrate this point are images that contain occlusions (**Figure 8A**, right): What sense does it make to ask about the ground truth if it could be anything behind this occlusion, and we will never be able to tell from the incomplete data in the image? It only makes sense to ask what it looks like to a particular observer, so by forcing models

to match the "ground truth," we may in fact be pushing them to solve the wrong problem.

Similarly, raters are subject to their semantic knowledge. A human figure in a yellow skirt (**Figure 8B**) might be annotated as a human figure rather a body and a skirt separately. But for a model lacking extensive semantic knowledge (or statistical co-occurrences of higher-level entities), there is no reason why that yellow blob that happens to be a skirt could not be an occluder, unrelated to the human (like a flying broomstick). Regardless of whether or not the model combines the two into a single object, it does not mean that the model performed an incorrect initial segmentation. Thus, one needs to be very careful when defining what a correct segmentation is for a given model. A ground truth for one model might not be a ground truth for another.

Perhaps due to the lack of the ground truth, object localization is usually treated as accurate if at least 50% of the box containing the object overlaps with the box proposed by the model (Russakovsky et al., 2013). While finding the bounding box can often provide a good first guess of an object's location, as discussed in Section "Feature Interpolation," it is clear that this measure is far from the explicit human knowledge of the precise boundary and location of an object (**Figure 8C**). As a result, a model that is performing well according to this benchmark might be doing so in a completely different way than we expect or want. For example, an interesting study by Landecker et al. (2013) attempted to track down which parts of an image end up being most important for classification in hierarchical networks. Curiously, they found that sometimes object classification decision was based on completely irrelevant information, such as a background whose statistics happened to match certain object characteristics (**Figure 8D**). Szegedy et al. (2013) provided another striking example where they showed that in a standard deep learning setup for every image it was possible to construct another perceptually indistinguishable image that would nevertheless be categorized incorrectly by the same network. Similarly, analyzing top-performing models in the Image Net Large Scale Visual Recognition Challenge 2012, Russakovsky et al. (2013) observed that while such models tend to provide rather accurate locations of detected objects, their performance deteriorates significantly with more objects or clutter. If object shapes were explicitly represented, clutter would play a much smaller role in localization errors. Finally, Torralba and Efros (2011) showed that models trained on one dataset often perform poorly on another dataset for the same categories of objects. What these models are learning then remains rather questionable. (However, note that there are also examples of models that are capable of generalizing across datasets; see Razavian et al., 2014.)

Finally, a model's output is extremely context dependent. For example, imagine that you are presented with a screen with one stimulus at a top and three below, as in **Figure 8E**, left. You are asked to indicate which item at the bottom matches best the one at the top. Most people would probably choose "Q." But now imagine the stimuli were slightly changed (**Figure 8E**, right). Most people would now go for "X." But how would a model know that? It should somehow take it into account that the colors of "O" and "X" match while "T" and "Q" have some other colors and it should also know that color is more important to the visual system than shape. In other words, it needs a lot of basic knowledge, or basic reasoning skills, that are arguably even harder to build in the system than vision itself.

To avoid some of the listed problems, we suggest using artificially generated scenes, such as the one in **Figure 4**. They can be rendered to contain many difficult features that are abundant in natural images, including shadows, occlusions, clutter, and realistic textures. However, unlike natural images, such scenes do have a well-defined ground truth because they are rendered from three dimensional models. Moreover, since they lack known objects, a good model should be completely capable of dividing an image into surfaces all on its own with little or no mistakes. If the model fails, it is a clear indication that intermediate representations are not being constructed properly yet.

Another possibility to evaluate model's performance is to use the extracted statistics to synthesize new images. This approach was taken by Portilla and Simoncelli (2000) who convincingly showed that their texture synthesis model was accurate by presenting an original texture and synthetically generated ones using the computed statistics. Arguably, such approach would be much trickier to implement for a synthesis of objects (Portilla and Simoncelli's procedure fails to produce coherent objects) but then the model's performance would be more directly observable and would point to issues where the algorithm needs an improvement.

LIMITATIONS AND CONCLUSION

In this paper, we provided a synthesis of the classical works in psychology and recent advances in visual neuroscience and computer vision into a single unified framework of mid-level computations. We hypothesized that two basic mechanisms, namely, similarity estimation and pooling, implemented hierarchically and reiterated via recurrent processing, appear to be sufficient to account for the computational goals of mid-level vision and the available empirical data.

Admittedly, many details in the proposed framework remain speculative at this point. While we provided the sketch of each processing stage (including the initial feature extraction, junction and curvature computation, region growing, border-ownership assignment, and figure-ground organization), it remains to be seen to what extent these computations are robust in natural image processing. Similarly, while the framework can flexibly operate in various feature spaces, we do not propose which features in particular should be included and how different cues could be combined. Learning the weights of these cues is crucial if we want the proposed framework to apply for real images. One possibility is that the proposed computations can be implemented in the standard deep learning networks (by replacing nonlinearity and normalization steps with similarity estimation, and also performing feature inference instead of a simple filtering).

Another possibility, given that, unlike deep networks, the proposed architecture does not require semantic knowledge to be trained, observing certain feature co-occurrences (see Geisler, 2008, for a review) would be a simpler way to learn and adjust these weights. Even more powerful cues would be available from dynamic or stereo-defined inputs, given their tremendous role in bootstrapping the visual system (Ostrovsky et al., 2006, 2009)

Furthermore, we restricted the scope of our discussions to the construction of the initial figure-ground organization briefly after stimulus onset. This choice has been motivated by our interest to advance the idea that image segmentation and figure-ground organization might be rapid, nearly feedforward computations. However, recurrent processing loops are undoubtedly necessary to improve the constructed surfaces and meet task demands. We considered several alternatives for such computations in Section "Evaluating Performance," but the details of such top-down refinement remain to be worked out.

More than anything, this paper is our manifesto on the importance of intermediate computations. We are calling for a reconsideration of the role of mid-level vision and propose that implementing several basic mechanisms might provide an significant step forward in understanding the functioning of primate visual system.

ACKNOWLEDGMENTS

This work was supported in part by a Methusalem Grant (METH/08/02) awarded to Johan Wagemans from the Flemish Government. Jonas Kubilius is a research assistant of the Research Foundation—Flanders (FWO). We thank Naoki Kogo, Bart Machilsen, Lee de-Wit, Pieter Roelfsema, James Elder, Pieter Moors, Maarten Demeyer, and the reviewers of this paper for fruitful discussions and criticism, and Tom Putzeys for generating 3D scenes.

REFERENCES

Σ64. (2012). *The Forest in Yakushima, Kagoshima Pref., Japan*. Available online at: http://commons.wikimedia.org/wiki/File:Forest_in_Yakushima_08.jpg (Accessed November 26, 2014).

Allen, H. A., Humphreys, G. W., Colin, J., and Neumann, H. (2009). Ventral extrastriate cortical areas are required for human visual texture segmentation. *J. Vis.* 9:2. doi: 10.1167/9.9.2

Amir, O., Biederman, I., and Hayworth, K. J. (2012). Sensitivity to nonaccidental properties across various shape dimensions. *Vision Res.* 62, 35–43. doi: 10.1016/j.visres.2012.03.020

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupé, J.-M., Bullier, J., and Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *J. Neurosci.* 22, 8633–8646.

Arall, M., Romeo, A., and Supèr, H. (2012). "Role of feedforward and feedback projections in figure-ground responses," in *Visual Cortex—Current Status and Perspectives*, ed S. Molotchnikoff (InTech). Available online at: http://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/role-of-feedforward-and-feedback-projections-in-figure-ground-responses

Arbeláez, P., Maire, M., Fowlkes, C., and Malik, J. (2011). Contour detection and hierarchical image segmentation. *Pattern Anal. Mach. Intell. IEEE Trans.* 33, 898–916. doi: 10.1109/TPAMI.2010.161

Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., and Malik, J. (2014). "Multiscale combinatorial grouping," in *Computer Vision and Pattern Recognition*. Available online at: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/mcg/

Attneave, F. (1954). Some informational aspects of visual perception. *Psychol. Rev.* 61, 183–193. doi: 10.1037/h0054663

Balas, B., Nakano, L., and Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* 9:13. doi: 10.1167/9.12.13

Ban, H., Yamamoto, H., Hanakawa, T., Urayama, S., Aso, T., Fukuyama, H., et al. (2013). Topographic representation of an occluded object and the effects of spatiotemporal context in human early visual areas. *J. Neurosci.* 33, 16992–17007. doi: 10.1523/JNEUROSCI.1455-12.2013

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629. doi: 10.1038/nrn1476

Barenholtz, E., and Tarr, M. J. (2007). "Reconsidering the role of structure in vision," in *Categories in Use*, Vol. 47, eds A. Markman and B. Ross (San Diego, CA: Academic Press), 157–180.

Barlow, H. B. (1961). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication*, ed W. Rosenblith (Cambridge, MA: MIT Press), 217–234.

bengt-re. (2009). *Car 132*. Available online at: https://www.flickr.com/photos/bengt-re/3889978926/ (Accessed November 26, 2014).

Berbezier, M. (2008). *IMG_3536*. Available online at: https://www.flickr.com/photos/berbezier/2399217350/ (Accessed November 26, 2014).

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Blum, H. (1973). Biological shape and visual science (part I). *J. Theor. Biol.* 38, 205–287. doi: 10.1016/0022-5193(73)90175-6

Bosch, A., Zisserman, A., and Munoz, X. (2007). "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (New York, NY).

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010). "Learning mid-level features for recognition," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (San Francisco, CA).

Brincat, S. L., and Connor, C. E. (2006). Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49, 17–24. doi: 10.1016/j.neuron.2005.11.026

Bullier, J. (2001). Integrated model of visual processing. *Brain Res. Rev.* 36, 96–107. doi: 10.1016/S0165-0173(01)00085-6

Bushnell, B. N., Harding, P. J., Kosai, Y., and Pasupathy, A. (2011). Partial occlusion modulates contour-based shape encoding in primate area V4. *J. Neurosci.* 31, 4012–4024. doi: 10.1523/JNEUROSCI.4766-10.2011

Camera Eye Photography. (2013). *Trio, Quartet, or Quintet of Objects... Week #26 [26 of 52]*. Available online at: https://www.flickr.com/photos/camera_is_a_mirror_with_memory/9126653341/ (Accessed November 26, 2014).

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597. doi: 10.1523/JNEUROSCI.3726-05.2005

Chen, L. (1982). Topological structure in visual perception. *Science* 218, 699–700. doi: 10.1126/science.7134969

Chen, L. (2005). The topological approach to perceptual organization. *Vis. Cogn.* 12, 553–637. doi: 10.1080/13506280444000256

Choi, M. J., Torralba, A., and Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognit. Lett.* 33, 853–862. doi: 10.1016/j.patrec.2011.12.004

Coates, A., Karpathy, A., and Ng, A. (2012). "Emergence of object-selective features in unsupervised feature learning," *Presented at the Advances in Neural Information Processing Systems 25*. Available online at: http://nips.cc/Conferences/2012/Program/event.php?ID=3342

Cox, D. D. (2014). Do we understand high-level vision? *Curr. Opin. Neurobiol.* 25, 187–193. doi: 10.1016/j.conb.2014.01.016

Craft, E., Schütze, H., Niebur, E., and von der Heydt, R. (2007). A neural model of figure-ground organization. *J. Neurophysiol.* 97, 4310–4326. doi: 10.1152/jn.00203.2007

D'Antona, A. D., Perry, J. S., and Geisler, W. S. (2013). Humans make efficient use of natural image statistics when performing spatial interpolation. *J. Vis.* 13:11. doi: 10.1167/13.14.11

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, Vol. 1 (San Diego, CA), 886–893. doi: 10.1109/CVPR.2005.177

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (Miami, FL), 248–255. doi: 10.1109/CVPR.2009.5206848

Desimone, R., Albright, T., Gross, C., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4, 2051–2062.

DiCarlo, J. J., and Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341. doi: 10.1016/j.tics.2007.06.010

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

Dobbins, A., Zucker, S. W., and Cynader, M. S. (1987). Endstopped neurons in the visual cortex as a substrate for calculating curvature. *Nature* 329, 438–441. doi: 10.1038/329438a0

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science* 293, 2470–2473. doi: 10.1126/science.1063414

Elder, J. H., and Zucker, S. W. (1996). "Computing contour closure," in *Computer Vision—ECCV'96*, eds B. Buxton and R. Cipolla (Berlin Heidelberg: Springer), 399–412. Available online at: http://link.springer.com/chapter/10.1007/BFb0015553

El-Shamayleh, Y., and Movshon, J. A. (2011). Neuronal responses to texture-defined form in macaque visual area V2. *J. Neurosci.* 31, 8543–8555. doi: 10.1523/JNEUROSCI.5974-10.2011

Epstein, R., and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature* 392, 598–601. doi: 10.1038/33402

Feldman, J. (1997). Regularity-based perceptual grouping. *Comput. Intell.* 13, 582–623. doi: 10.1111/0824-7935.00052

Feldman, J. (2003). What is a visual object? *Trends Cogn. Sci.* 7, 252–256. doi: 10.1016/S1364-6613(03)00111-6

Felleman, D. J., and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47. doi: 10.1093/cercor/1.1.1

Felzenszwalb, P. F., and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *Int. J. Comput. Vis.* 59, 167–181. doi: 10.1023/B:VISI.0000022288.19776.77

Fidler, S., Boben, M., and Leonardis, A. (2009). "Learning hierarchical compositional representations of object structure," in *Object Categorization: Computer and Human Vision Perspectives*, *1st Edn.*, eds S. J. Dickinson, A. Leonardis, B. Schiele, and M. J. Tarr (New York, NY: Cambridge University Press), 196–215.

Freeman, J., and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1201. doi: 10.1038/nn.2889

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

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annu. Rev. Psychol.* 59, 167–192. doi: 10.1146/annurev.psych.58.110405.085632

Geisler, W. S., and Perry, J. S. (2009). Contour statistics in natural images: grouping across occlusions. *Vis. Neurosci.* 26, 109–121. doi: 10.1017/S0952523808080875

Geisler, W. S., and Super, B. J. (2000). Perceptual organization of two-dimensional patterns. *Psychol. Rev.* 107, 677–708. doi: 10.1037/0033-295X.107.4.677

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton, MI: Mifflin.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available online at: http://arxiv.org/abs/1311.2524

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Percept. Psychophys.* 55, 48–121. doi: 10.3758/BF03206880

Hansen, T., and Neumann, H. (2004). Neural Mechanisms for the Robust Representation of Junctions. *Neural Comput.* 16, 1013–1037. doi: 10.1162/089976604773135087

Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. (2014). "Simultaneous detection and segmentation," in *European Conference on Computer Vision (ECCV)*. Available online at: http://arxiv.org/abs/1407.1808

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hochstein, S., and Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron* 36, 791–804. doi: 10.1016/S0896-6273(02)01091-7

Hubel, D. H., and Wiesel, T. N. (1961). Integrative action in the cat's lateral geniculate body. *J. Physiol.* 155, 385–398.

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Hubel, D. H., and Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* 28, 229–289.

Hummel, J. E., and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517. doi: 10.1037/0033-295X.99.3.480

Hung, C.-C., Carlson, E. T., and Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron* 74, 1099–1113. doi: 10.1016/j.neuron.2012.04.029

Huth, A. G., Nishimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Hyvärinen, A., and Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.* 12, 1705–1720. doi: 10.1162/089976600300015312

Hyvärinen, A., and Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* 41, 2413–2423. doi: 10.1016/S0042-6989(01)00114-6

Ion, A., Carreira, J., and Sminchisescu, C. (2013). Probabilistic joint image segmentation and labeling by figure-ground composition. *Int. J. Comput. Vis.* doi: 10.1007/s11263-013-0663-7

Ito, M., and Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *J. Neurosci.* 24, 3313–3324. doi: 10.1523/JNEUROSCI.4364-03.2004

Jehee, J. F. M., Lamme, V. A. F., and Roelfsema, P. R. (2007). Boundary assignment in a recurrent network architecture. *Vision Res.* 47, 1153–1165. doi: 10.1016/j.visres.2006.12.018

Joo, S. J., and Murray, S. O. (2014). Contextual effects in human visual cortex depend on surface structure. *J. Neurophysiol.* 111, 1783–1791. doi: 10.1152/jn.00671.2013

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. *Riv. Psicol.* 49, 7–30.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311.

Kayaert, G., Biederman, I., Op de Beeck, H. P., and Vogels, R. (2005a). Tuning for shape dimensions in macaque inferior temporal cortex. *Eur. J. Neurosci.* 22, 212–224. doi: 10.1111/j.1460-9568.2005.04202.x

Kayaert, G., Biederman, I., and Vogels, R. (2005b). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cereb. Cortex* 15, 1308–1321. doi: 10.1093/cercor/bhi014

Kogo, N., and Froyen, V. (2014). "Border-ownership computation reflecting consistency of surface properties," *Presented at the Vision Sciences Society* (St. Pete Beach, FL).

Kogo, N., Strecha, C., Van Gool, L., and Wagemans, J. (2010). Surface construction by a 2-D differentiation–integration process: a neurocomputational model for perceived border ownership, depth, and lightness in Kanizsa figures. *Psychol. Rev.* 117, 406–439. doi: 10.1037/a0019076

Kogo, N., and van Ee, R. (2014). "Neural mechanisms of figure-ground organization: border-ownership, competition and perceptual switching," in *Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford, UK: Oxford University Press).

Kornblith, S., Cheng, X., Ohayon, S., and Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron* 79, 766–781. doi: 10.1016/j.neuron.2013.06.015

Kosai, Y., El-Shamayleh, Y., Fyall, A. M., and Pasupathy, A. (2014). The role of visual area V4 in the discrimination of partially occluded shapes. *J. Neurosci.* 34, 8570–8584. doi: 10.1523/JNEUROSCI.1375-14.2014

Kourtzi, Z., and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science* 293, 1506–1509. doi: 10.1126/science.1061133

Kreiman, G. (2013). "Computational models of visual object recognition," in *Principles of Neural Coding*, Vol. 1-0. (CRC Press), 565–580. Available online at: http://www.crcnetbase.com/doi/abs/10.1201/b14756-33 doi: 10.1201/b14756-33

Krizhevsky, A., Sutskever,I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc), 1097–1105. Available online

at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Kubilius, J., Wagemans, J., and Op de Beeck, H. P. (2014). Encoding of configural regularity in the human visual system. *J. Vis.* 14:11. doi: 10.1167/14.9.11

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.* 16, 37–68.

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* 15, 1605–1615.

Lamme, V. A. F., Rodriguez-Rodriguez, V., and Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cereb. Cortex* 9, 406–413. doi: 10.1093/cercor/9.4.406

Landecker, W., Thomure, M. D., Bettencourt, L. M. A., Mitchell, M., Kenyon, G. T., and Brumby, S. P. (2013). "Interpreting individual classifications of hierarchical networks," in *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (Singapore).

Larsson, J., Landy, M. S., and Heeger, D. J. (2006). Orientation-selective adaptation to first- and second-order patterns in human visual cortex. *J. Neurophysiol.* 95, 862–881. doi: 10.1152/jn.00668.2005

Layton, O. W., Mingolla, E., and Yazdanbakhsh, A. (2012). Dynamic coding of border-ownership in visual cortex. *J. Vis.* 12:8. doi: 10.1167/12.13.8

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551. doi: 10.1162/neco.1989.1.4.541

Lee, T. S., Mumford, D., and Yuille, A. (1992). "Texture segmentation by minimizing vector-valued energy functionals: the coupled-membrane model," in *Computer Vision—ECCV'92*, ed G. Sandini (Berlin Heidelberg: Springer), 165–173. Available online at: http://link.springer.com/chapter/10.1007/3-540-55426-2_19

Leibe, B., Leonardis, A., and Schiele, B. (2008). Robust object detection with interleaved categorization and segmentation. *Int. J. Comput. Vis.* 77, 259–289. doi: 10.1007/s11263-007-0095-3

Lescroart, M. D., Biederman, I., Yue, X., and Davidoff, J. (2010). A cross-cultural study of the representation of shape: sensitivity to generalized cone dimensions. *Visual Cogn.* 18, 50–66. doi: 10.1080/13506280802507806

Li, F. F., VanRullen, R., Koch, C., and Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9596–9601. doi: 10.1073/pnas.092277599

Li, G., Yao, Z., Wang, Z., Yuan, N., Talebi, V., Tan, J., et al. (2014). Form-cue invariant second-order neuronal responses to contrast modulation in primate area V2. *J. Neurosci.* 34, 12081–12092. doi: 10.1523/JNEUROSCI.0211-14.2014

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

Lu, C., and Tang, X. (2014). Surpassing human-level face verification performance on LFW with GaussianFace. *arXiv:1404.3840 [cs, Stat]*. Available online at: http://arxiv.org/abs/1404.3840

Marcar, V. L., Raiguel, S. E., Xiao, D., and Orban, G. A. (2000). Processing of kinetically defined boundaries in areas V1 and V2 of the macaque monkey. *J. Neurophysiol.* 84, 2786–2798.

Mareschal, I., and Baker, C. L. (1998). A cortical locus for the processing of contrast-defined contours. *Nat. Neurosci.* 1, 150–154. doi: 10.1038/401

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.

McDermott, J. (2004). Psychophysics with junctions in real images. *Perception* 33, 1101–1127. doi: 10.1068/p5265

Mihalas, S., Dong, Y., von der Heydt, R., and Niebur, E. (2011). Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7583–7588. doi: 10.1073/pnas.1014655108

Mumford, D., and Shah, J. (1989). Optimal approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.* 42, 577–685. doi: 10.1002/cpa.3160420503

Murray, R. F., Sekuler, A. B., and Bennett, P. J. (2001). Time course of amodal completion revealed by a shape discrimination task. *Psychon. Bull. Rev.* 8, 713–720. doi: 10.3758/BF03196208

Nakayama, K., He, Z. J., and Shimojo, S. (1995). "Visual surface representation: a critical link between lower-level and higher-level vision," in *Visual cognition: An invitation to cognitive science, 2nd Edn*, Vol. 2, eds S. M. Kosslyn and D. N. Osherson (Cambridge, MA: The MIT Press), 1–70. Available online at: http://visionlab.harvard.edu/members/ken/Papers/077NKHeShimojoMIT1995b.pdf

Nothdurft, H. C. (1994). Common properties of visual segmentation. *Ciba Found. Symp.* 184, 245–259. discussion: 260–271.

Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175. doi: 10.1023/A:1011139631724

Oliva, A., and Torralba, A. (2006). "Chapter 2 Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research*, Vol. 155, Part B, eds S. L. Macknik, L. M. Martinez, J. -M. Alonso, P. U. Tse and S. Martinez-Conde (Elsevier), 23–36. Available online at: http://www.sciencedirect.com/science/article/pii/S0079612306550022

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Olshausen, B. A., and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14, 481–487. doi: 10.1016/j.conb.2004.07.007

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* 4, 1244–1252. doi: 10.1038/nn767

Ostrovsky, Y., Andalman, A., and Sinha, P. (2006). Vision following extended congenital blindness. *Psychol. Sci.* 17, 1009–1014. doi: 10.1111/j.1467-9280.2006.01827.x

Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., and Sinha, P. (2009). Visual parsing after recovery from blindness. *Psychol. Sci.* 20, 1484–1491. doi: 10.1111/j.1467-9280.2009.02471.x

Pan, Y., Chen, M., Yin, J., An, X., Zhang, X., Lu, Y., et al. (2012). Equivalent representation of real and illusory contours in macaque V4. *J. Neurosci.* 32, 6760–6770. doi: 10.1523/JNEUROSCI.6140-11.2012

Paris, S., and Durand, F. (2007). "A topological approach to hierarchical segmentation using mean shift," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07* (Minneapolis, MN), 1–8.

Pasupathy, A., and Connor, C. E. (1999). Responses to contour features in macaque area V4. *J. Neurophysiol.* 82, 2490–2502.

Pasupathy, A., and Connor, C. E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* 86, 2505–2519.

Pasupathy, A., and Connor, C. E. (2002). Population coding of shape in area V4. *Nat. Neurosci.* 5, 1332–1338. doi: 10.1038/nn972

Peelen, M. V., and Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *J. Neurophysiol.* 93, 603–608. doi: 10.1152/jn.00513.2004

Peterson, M. A. (1994). Object recognition processes can and do operate before figure-ground organization. *Curr. Dir. Psychol. Sci.* 3, 105–111. doi: 10.1111/1467-8721.ep10770552

Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput. Biol.* 4:e27. doi: 10.1371/journal.pcbi.0040027

Poort, J., Raudies, F., Wannig, A., Lamme, V. A. F., Neumann, H., and Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron* 75, 143–156. doi: 10.1016/j.neuron.2012.04.032

Portilla, J., and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40, 49–70. doi: 10.1023/A:1026553619983

Potter, M. C. (1976). Short-term conceptual memory for pictures. *J. Exp. Psychol. Hum. Learn. Mem.* 2, 509–522. doi: 10.1037/0278-7393.2.5.509

Purves, D., Monson, B. B., Sundararajan, J., and Wojtach, W. T. (2014). How biological vision succeeds in the physical world. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4750–4755. doi: 10.1073/pnas.1311309111

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition* 80, 127–158. doi: 10.1016/S0010-0277(00)00156-6

Qiu, F. T., Sugihara, T., and von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, 10, 1492–1499. doi: 10.1038/nn1989

Quiroga, R. Q., Mukamel, R., Isham, E. A., Malach, R., and Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3599–3604. doi: 10.1073/pnas.0707043105

Ramenahalli, S., Mihalas, S., and Niebur, E. (2014). Local spectral anisotropy is a valid cue for figure-ground organization in natural scenes. *Vision Res.* 103, 116–126. doi: 10.1016/j.visres.2014.08.012

Ramsden, B. M., Hung, C. P., and Roe, A. W. (2001). Real and illusory contour processing in area V1 of the primate: a cortical balancing act. *Cereb. Cortex* 11, 648–665. doi: 10.1093/cercor/11.7.648

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Rauschenberger, R., Liu, T., Slotnick, S. D., and Yantis, S. (2006). Temporally unfolding neural representation of pictorial occlusion. *Psychol. Sci.* 17, 358–364. doi: 10.1111/j.1467-9280.2006.01711.x

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. *arXiv:1403.6382 [cs]*. Available online at: http://arxiv.org/abs/1403.6382

Regan, D. (2000). *Human Perception of Objects: Early Visual Processing of Spatial form Defined by Luminance, Color, Texture, Motion, and Binocular Disparity*. Sunderland, MA: Sinauer Associates.

Reza. (2009). *Mountains*. Available online at: https://www.flickr.com/photos/r-z/3690169484/ (Accessed November 26, 2014).

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Ringach, D. L., and Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Res.* 36, 3037–3050. doi: 10.1016/0042-6989(96)00062-4

Rodríguez-Sánchez, A. J., and Tsotsos, J. K. (2012). The roles of endstopped and curvature tuned computations in a hierarchical representation of 2D shape. *PLoS ONE* 7:e42058. doi: 10.1371/journal.pone.0042058

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* 29, 203–227. doi: 10.1146/annurev.neuro.29.051605.112939

Roelfsema, P. R., Lamme, V. A. F., and Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nat. Neurosci.* 7, 982–991. doi: 10.1038/nn1304

Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., and Bosch, H. (2002). Figure—ground segregation in a recurrent network architecture. *J. Cogn. Neurosci.* 14, 525–537. doi: 10.1162/08989290260045756

Roelfsema, P. R., van Ooyen, A., and Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends Cogn. Sci.* 14, 64–71. doi: 10.1016/j.tics.2009.11.005

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., and Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *J. Vis.* 12:14. doi: 10.1167/12.4.14

Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., and Wattenberg, M. (2009). "An intuitive model of perceptual grouping for HCI design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1331–1340. Available online at: http://dl.acm.org/citation.cfm?id=1518903

Roskies, A. L. (1999). The binding problem. *Neuron* 24, 7–9. doi: 10.1016/S0896-6273(00)80817-X

Russakovsky, O., Deng, J., Huang, Z., Berg, A. C., and Fei-Fei, L. (2013). "Detecting avocados to zucchinis: what have we done, and where are we going?" in *2013 IEEE International Conference on Computer Vision (ICCV)* (Sydney, NSW), 2064–2071. doi: 10.1109/ICCV.2013.258

Russell, A. F., Mihalaş, S., von der Heydt, R., Niebur, E., and Etienne-Cummings, R. (2014). A model of proto-object based saliency. *Vision Res.* 94, 1–15. doi: 10.1016/j.visres.2013.10.005

Scharstein, D., and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vis.* 47, 7–42. doi: 10.1023/A:1014573219977

Schmid, A. M., Purpura, K. P., and Victor, J. D. (2014). Responses to orientation discontinuities in V1 and V2: physiological dissociations and functional implications. *J. Neurosci.* 34, 3559–3578. doi: 10.1523/JNEUROSCI.2293-13.2014

Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Comput.* 4, 234–242. doi: 10.1162/neco.1992.4.2.234

Sekuler, A. B., and Palmer, S. E. (1992). Perception of partly occluded objects: a microgenetic analysis. *J. Exp. Psychol.* 121, 95–111. doi: 10.1037/0096-3445.121.1.95

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). OverFeat: integrated recognition, localization and detection using convolutional networks. *arXiv:1312.6229 [cs]*. Available online at: http://arxiv.org/abs/1312.6229

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Sharon, E., Galun, M., Sharon, D., Basri, R., and Brandt, A. (2006). Hierarchy and adaptivity in segmenting visual scenes. *Nature* 442, 810–813. doi: 10.1038/nature04977

Sheila in Moonducks. (2010). *Tree*. Available online at: https://www.flickr.com/photos/aspis7/5075169756/ (Accessed November 26, 2014).

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 888–905. doi: 10.1109/34.868688

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193

Singer, W., and Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 18, 555–586. doi: 10.1146/annurev.ne.18.030195.003011

Snowdog. (2005). *Leuven - Grote Markt*. Available online at: http://commons.wikimedia.org/wiki/File:Leuven-Grote-Markt.jpg (Accessed November 26, 2014).

Song, Y., and Baker, C. L. (2007). Neuronal response to texture- and contrast-defined boundaries in early visual cortex. *Visual Neurosci.* 24, 65–77. doi: 10.1017/S0952523807070113

Supèr, H., and Lamme, V. A. F. (2007). Altered figure-ground perception in monkeys with an extra-striate lesion. *Neuropsychologia* 45, 3329–3334. doi: 10.1016/j.neuropsychologia.2007.07.001

Supèr, H., Romeo, A., and Keil, M. (2010). Feed-forward segmentation of figure-ground and assignment of border-ownership. *PLoS ONE* 5:e10705. doi: 10.1371/journal.pone.0010705

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2014). Going deeper with convolutions. *arXiv:1409.4842 [cs]*. Available online at: http://arxiv.org/abs/1409.4842

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*. Available online at: http://arxiv.org/abs/1312.6199

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). "DeepFace: closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Available online at: https://www.facebook.com/publications/546316888800776/ doi: 10.1109/CVPR.2014.220

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Thielscher, A., and Neumann, H. (2003). Neural mechanisms of cortico–cortical interaction in texture boundary detection: a modeling approach. *Neuroscience* 122, 921–939. doi: 10.1016/j.neuroscience.2003.08.050

Thielscher, A., and Neumann, H. (2005). Neural mechanisms of human texture processing: texture boundary detection and visual search. *Spat. Vis.* 18, 227–257. doi: 10.1163/1568568053320594

Thielscher, A., and Neumann, H. (2008). Globally consistent depth sorting of overlapping 2D surfaces in a model using local recurrent interactions. *Biol. Cybern.* 98, 305–337. doi: 10.1007/s00422-008-0211-7

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Todd, J. T., Weismantel, E., and Kallie, C. S. (2014). On the relative detectability of configural properties. *J. Vis.* 14:18. doi: 10.1167/14.1.18

Torralba, A., and Efros, A. A. (2011). "Unbiased look at dataset bias," in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI), 1521–1528.

Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. *Network (Bristol, England)* 14, 391–412. doi: 10.1088/0954-898X/14/3/302

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science* 311, 670–674. doi: 10.1126/science.1119983

Tschechne, S., and Neumann, H. (2014). Hierarchical representation of shapes in visual cortex–from localized features to figural shape segregation. *Front. Comput. Neurosci.* 8:93. doi: 10.3389/fncom.2014.00093

Tse, P. U. (1999). Volume completion. *Cogn. Psychol.* 39, 37–68. doi: 10.1006/cogp.1999.0715

Ullman, S., and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 992–1006. doi: 10.1109/34.99234

van Lier, R., van der Helm, P., and Leeuwenberg, E. (1994). Integrating global and local aspects of visual occlusion. *Perception* 23, 883–903. doi: 10.1068/p230883

van Lier, R. (1999). Investigating global effects in visual occlusion: from a partly occluded square to the back of a tree-trunk. *Acta Psychol. (Amst.)* 102, 203–220.

Vilankar, K. P., Golden, J. R., Chandler, D. M., and Field, D. J. (2014). Local edge statistics provide information regarding occlusion and nonocclusion edges in natural scenes. *J. Vis.* 14:13. doi: 10.1167/14.9.13

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001* (Vol. 1, I–511–I–518) (Kauai, HI). doi: 10.1109/CVPR.2001.990517

Vogels, R., Biederman, I., Bar, M., and Lorincz, A. (2001). Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *J. Cogn. Neurosci.* 13, 444–453. doi: 10.1162/08989290152001871

von der Heydt, R., and Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *J. Neurosci.* 9, 1731–1748.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science* 224, 1260–1262. doi: 10.1126/science.6539501

von der Malsburg, C. (1981). *The Correlation Theory of Brain Function.* Departmental Technical Report, MPI. Available online at: http://cogprints.org/1380/

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012a). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bull.* 138, 1172–1217. doi: 10.1037/a0029333

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., et al. (2012b). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol. Bull.* 138, 1218–1252. doi: 10.1037/a0029334

Wagemans, J., Lamote, C., and Gool, L. V. (1997). Shape equivalence under perspective and projective transformations. *Psychon. Bull. Rev.* 4, 248–253. doi: 10.3758/BF03209401

Wagemans, J., Van Gool, L., Lamote, C., and Foster, D. H. (2000). Minimal information to determine affine shape equivalence. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 443–468. doi: 10.1037/0096-1523.26.2.443

Walther, D., and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Netw.* 19, 1395–1407. doi: 10.1016/j.neunet.2006.10.001

Wehr, M., and Laurent, G. (1996). Odour encoding by temporal sequences of firing in oscillating neural assemblies. *Nature* 384, 162–166. doi: 10.1038/384162a0

Weidenbacher, U., and Neumann, H. (2009). Extraction of surface-related features in a recurrent model of V1-V2 interactions. *PLoS ONE* 4:e5909. doi: 10.1371/journal.pone.0005909

Witkin, A. P., and Tenenbaum, J. M. (1983). On the role of structure in vision. *Hum. Mach. Vis.* 1, 481–543. doi: 10.1016/B978-0-12-084320-6.50022-0

Wyatte, D., Herd, S., Mingus, B., and O'Reilly, R. (2012). The role of competitive inhibition and top-down feedback in binding during object recognition. *Front. Psychol.* 3:182. doi: 10.3389/fpsyg.2012.00182

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., and Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* 11, 1352–1360. doi: 10.1038/nn.2202

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yu, C.-P., Samaras, D., and Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *J. Vis.* 14:4. doi: 10.1167/14.7.4

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron* 47, 143–153. doi: 10.1016/j.neuron.2005.04.005

Zhou, H., Friedman, H. S., and von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20, 6594–6611.

Zucker, S. W. (2014). "Border inference and border ownership. The challenge of integrating geometry and topology," in *Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford, UK: Oxford University Press). doi: 10.1093/oxfordhb/9780199686858.013.020

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 May 2014; accepted: 17 November 2014; published online: 12 December 2014.

Citation: Kubilius J, Wagemans J and Op de Beeck HP (2014) A conceptual framework of computations in mid-level vision. Front. Comput. Neurosci. ***8****:158. doi: 10.3389/fncom.2014.00158*

This article was submitted to the journal Frontiers in Computational Neuroscience.