Recognition of Occluded Objects

Hanlin Tang and Gabriel Kreiman

Pattern recognition involves building a mental model to interpret incoming inputs. This incoming information is often incomplete and the mental model must extrapolate to complete the patterns, a process that is constrained by the statistical regularities in nature. Examples of pattern completion involve identification of objects presented under unfavorable luminance or interpretation of speech corrupted by acoustic noise. Pattern completion is also at the heart of other high-level cognitive phenomena including our ability to discern actions from still images or to predict behavioral patterns from observations.

Pattern completion constitutes a ubiquitous challenge during natural vision. Stimuli are often partially occluded, or degraded by changes in illumination and contrast. While much progress has been made towards understanding the mechanisms underlying recognition of complete objects, the neural computations underlying more challenging recognition problems such as object occlusion remain poorly understood. Here we generically refer to object occlusion to include any transformation where only partial information about an object is accessible (such as the multiple examples illustrated in Fig. 1), and refer to object completion as the ability to infer object identity from partial information (without necessarily implying that subjects perceptually fill in the missing information).

Understanding how the neural representations of visual signals are modified when objects are occluded is critical to developing biologically constrained computational models of occluded object recognition and may also shed light on how to solve manifestations of pattern completion in other domains. The development of feed-forward models for visual recognition of whole objects has been driven by behavioral and physiological experiments establishing the hierarchy of feature tuning and robustness to image transformations. Similarly, by systematically examining when and where neural representations that are robust to occlusion

H. Tang · G. Kreiman (✉)
Harvard Medical School, 1 Blackfan Circle, Karp 11217, Boston, MA 02115, USA
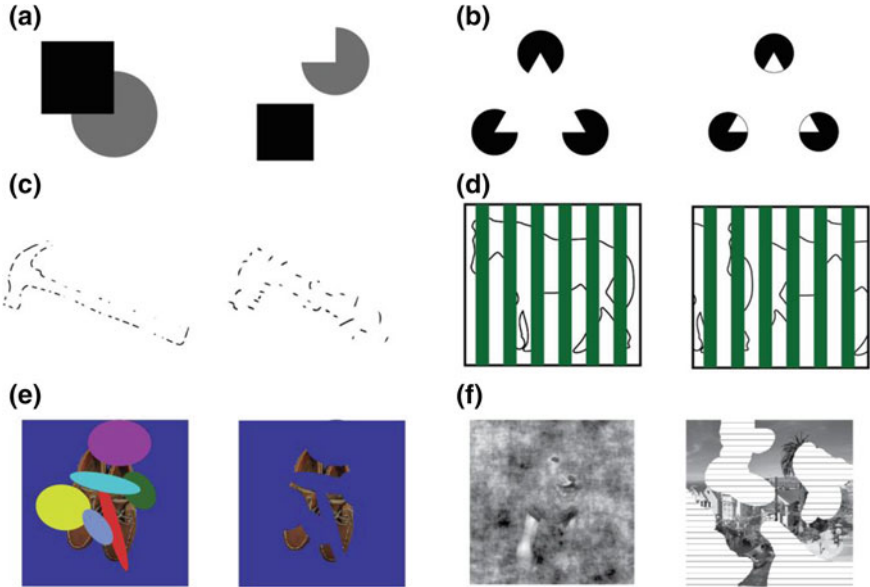e-mail: gabriel.kreiman@tch.harvard.edu

Fig. 1 Object completion examples **a** Occluded geometric shape (*left*) and its mosaic counterpart (*right*) (e.g. Murray 2004). **b** Example of modal completion inducing an illusory triangle (*left*). This percept is disrupted by adding edges to the inducers (*right*). **c** Line drawing of an object defined by disconnected segments and its fragmented counterpart, similar to (e.g. Doniger et al. 2000; Sehatpour et al. 2008). **d** Line drawing of an occluded object and its scrambled counterpart (e.g. Lerner et al. 2004, 2002). **e** Occluded object and its 'deleted' counterpart (e.g. Johnson and Olshausen 2005). **f** Example partial image of an object seen through bubbles with a phase-scrambled background to equalize contrast (e.g. Tang et al. 2014). **g** Example partial image of a scene (e.g. Nielsen et al. 2006)

emerge can help extend our theoretical understanding of vision and develop the next generation of computational models in vision. Understanding pattern completion and recognition of occluded objects is a challenging task: performance depends on the stimulus complexity, the type and amount of occlusion and the task itself. In this chapter, we summarize recent efforts to examine the mechanisms underlying recognition of partially occluded objects and discuss several avenues for future work towards a theory of object completion.

1 Visual System Hierarchy

Object recognition is orchestrated through a semi-hierarchical series of processing areas along ventral visual cortex (Connor et al. 2007; DiCarlo et al. 2012; Felleman and Van Essen 1991; Logothetis and Sheinberg 1996; Riesenhuber and Poggio 1999; Schmolesky et al. 1998; Tanaka 1996). At each step in this hierarchy, the

41 feature specificity of the neurons increases in complexity. For example, neurons in
42 primary visual cortex (V1), respond selectively to bars of a particular orientation
43 (Hubel and Wiesel 1959) whereas neurons in inferior temporal cortex respond
44 preferentially to complex shapes including faces and other objects (Desimone et al.
45 1984; Gross et al. 1969; Perrett et al. 1992; Richmond et al. 1983; Rolls 1991). In
46 addition to this increase in feature complexity, there is a concomitant progression in
47 the degree of tolerance to object transformations such as changes in object position
48 or scale (Hung et al. 2005; Ito et al. 1995; Logothetis et al. 1995). The selective and
49 tolerant physiological responses characterized in the macaque inferior temporal
50 cortex have also been observed in the human inferior temporal cortex (Allison et al.
51 1999; Liu et al. 2009). The timing of these neural responses provides important
52 constraints on the number of possible computations involved in visual recognition.
53 Multiple lines of evidence from human psychophysical measurements (Potter and
54 Levy 1969; Thorpe et al. 1996), macaque single unit recordings (Hung et al. 2005;
55 Keysers et al. 2001), human EEG (Thorpe et al. 1996) and human intracranial
56 recordings (Allison et al. 1999; Liu et al. 2009) have established that selective
57 responses to whole objects emerge within 100–150 ms of stimulus onset in the
58 highest echelons of the ventral visual stream.

59 Research over the last several decades characterizing the spatiotemporal
60 dynamics involved in the neural representation of objects in these successive areas
61 has led to the development of a theoretical framework to explain the mechanisms
62 underlying object recognition. An influential theoretical framework suggests that, to
63 a first approximation, processing of visual information traverses through the ventral
64 stream in a feed-forward fashion, without significant contributions from long
65 top-down feedback loops or within-area recurrent computations (Deco and Rolls
66 2004; Fukushima 1980; LeCun et al. 1998; Mel 1997; Olshausen et al. 1993;
67 Riesenhuber and Poggio 1999; Wallis and Rolls 1997). Consistent with this notion,
68 computational models of object recognition instantiating feed-forward processing
69 provide a parsimonious explanation for the selectivity and tolerances observed
70 experimentally (Serre et al. 2007b). The activity of these computational units at
71 various stages of processing captures the variability in the neural representation
72 from macaque single unit recordings along the visual hierarchy (Cadieu et al. 2014;
73 Yamins et al. 2014). These feed-forward computational models have inspired the
74 development of deep convolutional networks that demonstrate a significant degree
75 of success in a variety of computer vision approaches to object recognition (e.g.
76 Hinton and Salakhutdinov 2006; LeCun et al. 1998; Russakovsky et al. 2015; Sun
77 et al. 2014; Taigman et al. 2014).

78 These purely feedforward architectures do not incorporate any feedback or
79 recurrent connections. However, at the anatomical level, feedback and recurrent
80 connections figure prominently throughout the visual system (Felleman and Van
81 Essen 1991). In fact, quantitative anatomical studies have suggested that feedback
82 and recurrent connections significantly outnumber feedforward ones (Callaway
83 2004; Douglas and Martin 2004). The computational contributions of these feed-
84 back and recurrent projections are largely underexplored in existing computational
85 models of visual recognition because their underlying roles remain unclear. Several

investigators have suggested that these feedback and recurrent projections could play an important role during object recognition under conditions where the visual cues are impoverished (e.g. poor illumination, low contrast) or even partially missing (e.g. visual occlusion) (Carpenter and Grossberg 2002; Hopfield 1982; Mumford 1992; Tang et al. 2014; Wyatte et al. 2012a).

Following the approach suggested by Marr in his classic book on vision (Marr 1982), we subdivide our discussion of recognition of occluded objects into three parts: (i) definition of the *computational problem* by describing behavioral performance during recognition of occluded objects, (ii) characterization of the *implementation at the physical level* describing the neural responses to occluded objects and (iii) initial sketches of *theoretical ideas instantiated into computational models* that aim to recognize occluded objects.

2 The Computational Problem of Object Completion

Figure 1 shows examples of several images that induce object completion. In the natural world, objects can be partially occluded in multiple different ways due to the presence of explicit occluders, shadows, camouflage and differential illumination. Object completion is an ill-posed problem: in general, there are infinite ways of completing contours from partial information. The visual system must be able to infer what the object is despite the existence of all of these possible solutions consistent with the visual input.

2.1 Amodal Completion

Occluded shapes can be perceived as whole (Fig. 1a, compare left and right panels). Object completion can be *amodal* when there is an explicit occluder and the subject cannot see the contours behind the occluder despite being aware of the overall shape (Singh 2004). In contrast, in the famous illusory triangle example (Fig. 1b), Kanizsa describes the phenomenon known as *modal completion* whereby the object is completed by inducing illusory contours that are perceived by the observer (Kanizsa 1979). Because these illusory inducers are rare in natural vision, in this chapter we focus on amodal completion. Even though occluded or partial objects such as the ones shown in Fig. 1c, d are segmented, observers view the object as a single percept, not as disjointed segments. Amodal completion is also important for achieving this single `gestalt'. Investigators have used a variety of different stimuli to probe the workings of object completion, ranging from simple lines and geometric shapes to naturalistic objects such as the ones shown in Fig. 1e–g.

Psychophysical studies of amodal completion have provided many clues to the underlying computations (Kellman et al. 2001; Sekuler and Murray 2001). Amodal completion relies on an inferred depth between the occluder shape and the occluded
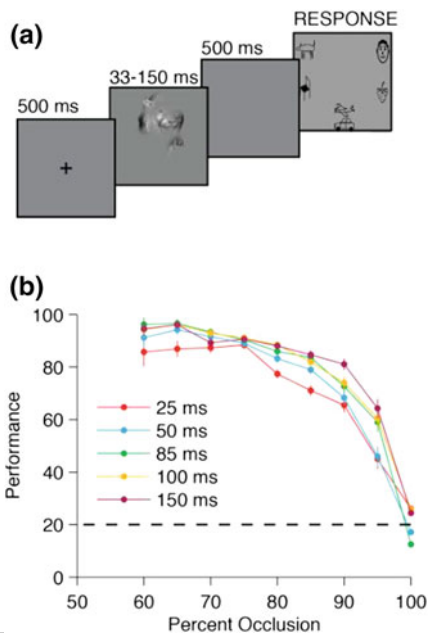
object, which in turns generates a surface-based representation of the scene (Nakayama et al. 1995). In fact, presence of the occluder aids in identifying the occluded object, as powerfully illustrated by the Bregman's occluded B letters (Bregman 1981) Grouping of different parts into a complete whole, and the 'completion' of missing lines and contours, represent an important component of object recognition. There are infinite possible ways of completing occluded objects. The ambiguities arise from the many combinations with which occluded edges, called `inducers' can be paired together, as well as the infinite possible contours between two pairs of inducers (Kellman et al. 2001; Nakayama et al. 1995; Ullman 1976). Despite the many possible solutions, the visual system typically arrives at a single (and correct) interpretation of the image.

The temporal dynamics of shape completion can constrain the computational steps involved in processing occluded images. Psychophysics experiments have measured the time course of amodal completion with a diverse array of experimental paradigms. The most common method is a contrast of an occluded shape against its mosaic parts (e.g. Figure 1a). For example, in the prime matching paradigm, subjects are first primed with a stimulus, and then asked to judge whether a pair of test stimuli represent the same or a different shape. Subjects are faster to correctly respond 'same' when the primed shape is the same as the test stimuli. When partly occluded objects are used as the prime, this priming effect depends on the exposure time (Sekuler and Palmer 1992). At short durations (50 ms), occluded objects primed subject's responses towards mosaic shapes, suggesting that 50 ms is not enough time for amodal completion of the prime stimulus. At longer durations (100 ms or more), the priming effect switched to favor whole shapes. Therefore, the authors estimate that amodal completion for simple geometric shapes takes between 100 and 200 ms, depending on the amount of occlusion (Sekuler et al. 1994). A different set of behavioral experiments suggests approximately the same time scales for amodal completion: in several studies, subjects are asked to discriminate shapes in a timed forced-choice task. Response times to occluded shapes lagged those to whole shapes by about 75–150 ms (Murray et al. 2001; Shore and Enns 1997). For naturalistic objects such as faces, however, Chen et al. report that amodal completion takes longer than 200 ms (Chen et al. 2009), well beyond the previous estimates based on simpler shapes and tasks.

2.2 From Amodal Completion to Recognition of Occluded Objects

Many studies of amodal completion have used simple shapes and contours, as outlined in the previous section. When recognizing whole objects, psychophysical studies using simple shapes and neurophysiological studies describing the linear filters in primary visual areas led to the basic building blocks to develop deep models of visual recognition capable of detecting complex shapes. Inspired by the

Fig. 2 Robust object categorization despite strong occlusion **a** Experiment timeline. Partial images similar to the ones shown in Fig. 1f were presented during exposure times ranging from 33 to 150 ms. Subjects performed a five alternative forced-choice categorization task determining whether the image contained animals, chairs, faces, fruits, or vehicles. **b** Performance as a function of percentage occlusion across n = 14 subjects for various exposure times. Error bars denote SEM. Dashed line indicates chance performance

¹⁶³ success of this approach, we assume that contour completion is one of the initial
¹⁶⁴ steps involved in interpreting complex objects that are partially occluded.

¹⁶⁵   Recognition remains robust to partial occlusion for complex objects. We used
¹⁶⁶ naturalistic objects that were occluded by presenting information through "bubbles"
¹⁶⁷ (Gosselin and Schyns 2001) (Fig. 1f). After a variable exposure time from 33 to
¹⁶⁸ 150 ms, subjects performed a five alternative forced-choice categorization task.
¹⁶⁹ Recognition was robust even when 80–90 % of the object was occluded across the
¹⁷⁰ various exposure times (Fig. 2). Similar results were obtained by (Wyatte et al.
¹⁷¹ 2012a). As illustrated by Bregman (1981), the presence of an occluder during object
¹⁷² completion aids recognition performance. For example, one study presented natural
¹⁷³ objects that were either occluded (Fig. 1e, left) or where the same object parts were
¹⁷⁴ deleted (Fig. 1e, right). Recognition was significantly impaired in the deleted part
¹⁷⁵ case compared to the occluded part case, but only when using high percentage of
¹⁷⁶ occlusion (>75 % missing pixels) (Johnson and Olshausen 2005).

¹⁷⁷   Intuitively, we would expect successful recognition to depend on the exact
¹⁷⁸ features shown. This intuition was quantitatively measured by (Gosselin and
¹⁷⁹ Schyns 2001): the facial features critical to recognition varied depending on the task
¹⁸⁰ (e.g. the eyes and eyebrows were more relevant for gender discrimination and the
¹⁸¹ mouth provided more information when evaluating expressiveness). Similar con-
¹⁸² clusions were reached when using the same paradigm to evaluate recognition in
¹⁸³ monkeys (Nielsen et al. 2006).

3 Neural Representation of Occluded Objects

A series of scalp electroencephalography (EEG) studies have measured the latency at which responses differ between occluded objects and suitable control images. Using simple geometric stimuli, differences between occluded shapes and notched shapes emerged at 140–240 ms (Murray 2004). Using more naturalistic stimuli (e.g. Figure 1e), other investigators report differential activity in the 130–220 ms (Chen et al. 2010) and 150–200 ms (Johnson and Olshausen 2005) ranges. In a more difficult task with fragmented line drawings that are progressively completed, Doniger et al. reported that differences are only observed in the 200–250 ms response window. Even though these studies use different stimuli and make different comparisons, they consistently conclude that amodal completion effects manifest within 130–250 ms.

Several neuroimaging (Hegde et al. 2008; Komatsu 2006; Lerner et al. 2004, 2002; Olson et al. 2004; Rauschenberger et al. 2004) and scalp EEG (Chen et al. 2010; Doniger et al. 2000; Johnson and Olshausen 2005) studies with more complex objects have contrasted activity changes between an occluded object and an appropriately scrambled counterpart (e.g. Fig. 1c, d). In these stimuli, the low-level features are maintained but disruption in their geometric arrangement destroys the percept. For example, investigators have reported differential activity in the lateral-occipital complex between occluded line drawings and their scrambled counterparts (Lerner et al. 2002). The authors reason that, since the occluded images elicit a larger response in the lateral-occipital complex (LOC) than scrambled images, the LOC could be involved in object completion. It should be noted that LOC also demonstrates increased activity to whole objects compared to scrambled versions of those objects without any occlusion (Grill-Spector et al. 2001). Thus, the increased responses to whole objects may not be necessarily related to object completion mechanisms per se, but rather neural activity related to perceptual recognition.

Similarly, EEG and intracranial studies compared line drawings against their fragmented counterparts (Doniger et al. 2000; Sehatpour et al. 2008) to measure the timing and brain regions involved in object completion. Sehatpour et al. worked with epilepsy patients who have intracranial electrodes implanted for clinical purposes. This approach allows a rare opportunity to record directly from human brain. The authors take advantage of simultaneous recordings from multiple brain regions to show that line fragments elicited greater coherence in the LOC-Prefrontal Cortex-Hippocampus network compared to scrambled line fragments. They suggest that this network synchrony is responsible for the perceptual line closure of objects.

In order to understand the neural mechanisms orchestrating object completion, it is also critical to examine the neural architectures that could implement the computational solutions suggested in the previous section. Essential aspects of object completion can be traced back to the earliest stages in visual processing. An early study demonstrated that neurons in area V2 show selective responses to illusory contours (Peterhans and von der Heydt 1991; von der Heydt et al. 1984). Other

227 work has demonstrated that even V1 neurons can respond to occluded shapes. One
228 study recorded single cells in macaque V1 when presented with occluded moving
229 bars (Sugita 1999). Approximately 12 % of orientation-selective cells responded to
230 the moving oriented bar even when it was occluded, thus potentially underlying the
231 phenomenology of amodal completion. These cells responded strongly only when
232 the occluder was presented in front of the moving bar (positive disparity), and not at
233 zero or negative disparity. Notably, responses to the occluded bar were not different
234 from those obtained when presenting the bar alone. These results have led to the
235 suggestion that amodal completion is achieved by contextual modulation from
236 outside the classical receptive field. While other studies have suggested that con-
237 textual modulation occurs with a delay of 50–70 ms with respect to the onset of the
238 visually evoked responses (Bakin et al. 2000; Zipser et al. 1996), Sugita did not
239 observe any latency delays for the amodally-completed response. The author sug-
240 gests that these contextual modulations may come from lateral connections or
241 feedback from proximal areas. In another study, responses to illusory contours in
242 V1 were delayed by about 55 ms compared to the response to real contours (Lee
243 and Nguyen 2001). Importantly, illusory contour responses appeared first in V2
244 before emerging in V1, suggesting that modal completion in V1 might require
245 feedback modulation from V2. Complementing these studies, psychophysical
246 studies on the effect of inferred depth and apparent motion on the perception of
247 occluded surfaces also conclude that amodal completion effects manifests in early
248 visual processing (Shimojo and Nakayama 1990a, b).

249 These neurophysiology studies have focused on the occlusion or inducing of
250 linear contours, where the inducers are close in proximity to the classical receptive
251 field. However in natural vision we complete curvilinear contours over distances
252 much longer than the width of classical V1 receptive fields. Often in these cases,
253 correct completion of an object depends on the global context in which the object is
254 embedded. Future studies are needed to examine whether and when V1 neurons
255 respond to completed contours of varying curvature, length, and context. **AQ1**

256 As outlined above, V1 neurons feed onto a cascade of semi-hierarchical pro-
257 cessing steps through V2 and V4, culminating in the inferior temporal cortex
258 (ITC) (Felleman and Van Essen 1991). How do these higher visual areas respond to
259 occluded shapes? Few studies have examined the responses in intermediate visual
260 areas to occluded shapes. A recent elegant study has begun to fill in this gap by
261 characterizing how macaque V4 neurons respond to different curvatures when they
262 are partially occluded by dots (Kosai et al. 2014). The authors report that neurons
263 can maintain selectivity within a range of occlusion. While the response onset times
264 of these neurons were not delayed by occlusion, the latency at which selectivity
265 arose was delayed by hundreds of milliseconds.

266 Kovacs et al. found that visually selective responses to complex shapes in ITC
267 were similar between whole shapes and occluded shapes defined by adding noise,
268 occluders or deleting shape parts (Kovács et al. 1995). Although selectivity to
269 complex shapes was retained despite up to ~50 % occlusion, the absolute magni-
270 tude of the responses was modulated linearly with the amount of occlusion. Con-
271 trary to what Kosai et al. found in V4, these authors observed delays of up to 50 ms

272 in the response latency of occluded shapes. While it is tempting to attribute this
273 discrepancy to differences in processing between V4 and IT, we note that the
274 stimuli and occluding patterns used are different between the two studies.

275 Nielsen et al. extended this work by examining the responses of IT neurons to
276 occluded objects embedded in naturalistic stimuli (Fig. 1f) (Nielsen et al. 2006a).
277 Using the bubbles paradigm (Gosselin and Schyns 2001), the authors defined parts
278 of an image that provided more diagnostic value (i.e. provided information that
279 aided recognition) versus other non-diagnostic parts. The authors first demonstrated
280 that monkeys and humans show striking behavioral similarities in terms of what
281 object parts are considered diagnostic (Nielsen et al. 2006b). For occluded scenes
282 containing diagnostic parts, both firing rates and local field potentials in ITC
283 remained largely invariant to significant amounts of occlusion, in contrast to the
284 findings of the Kovacs study with simpler stimuli (Kovacs et al. 1995). However,
285 for scenes that contained only non-diagnostic parts, the results from the Kovacs
286 study were reproduced—the firing rate varied linearly with the amount of occlusion.
287 This comparison also serves as a cautionary tale against extrapolating results based
288 on geometric shapes to the processing of more naturalistic stimuli because the
289 details of which features are revealed can play a very important role in dictating the
290 effects of occlusion. Issa et al. reached similar conclusions when demonstrating that
291 ITC responses selective to faces were particularly sensitive to occlusion of certain
292 parts (one eye) and that those parts could drive the responses almost as well as the
293 whole face (Issa and Dicarlo 2012). These results suggest that the robustness of the
294 neural representation to missing parts depends on the diagnosticity of the visible
295 features.

296 Tang et al. used intracranial recordings to evaluate how and when visually
297 selective responses to occluded objects emerge. Naturalistic objects were presented
298 through the bubbles paradigm (Gosselin and Schyns 2001) in a task similar to the
299 one illustrated in Fig. 2. The use of objects seen through bubbles evaluates the core
300 ability to spatially integrate multiple parts to subserve recognition. Figure 3 (left)
301 shows the responses from an example electrode in the fusiform gyrus that displayed
302 a strong response to face stimuli. Remarkably, the electrode showed similar
303 responses to images that displayed only ∼11 % of the object content (Fig. 3, col-
304 umns 2–6). Even in cases where the images shared essentially no common pixel
305 (e.g. columns 3 and 5), the responses remained similar. Overall, the magnitude of
306 the responses did not vary with the amount of visible pixels. Yet, the responses to
307 occluded shapes were not identical to those obtained upon presenting whole ima-
308 ges. A notable difference was that the responses to occluded objects were signifi-
309 cantly delayed (compare the position of the arrows in Fig. 3). Selective neural
310 response emerged with a delay of ∼100 ms. These delays were also apparent when
311 using a machine learning approach to decode the category or identity of the objects
312 from the physiological responses to the whole or occluded images (Fig. 4).

313 Image processing does not end with visual cortex. Information from visual
314 cortex is conveyed to frontal cortex and to medial temporal lobe structures
315 including the amygdala, hippocampus and entorhinal cortex. One group recorded
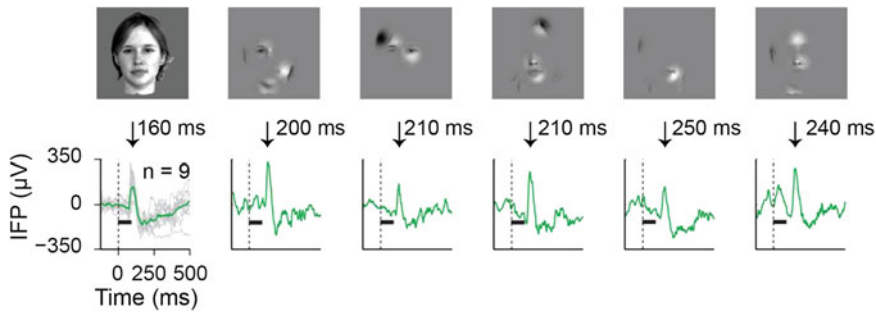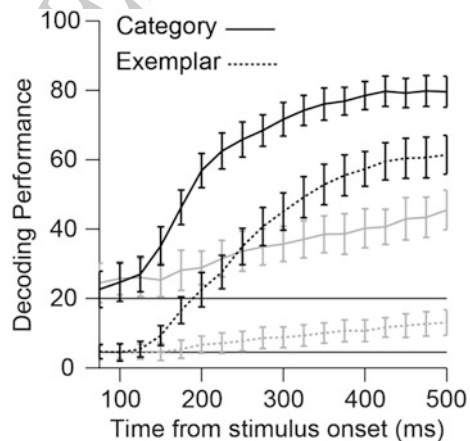316 single unit activity in the amygdala of human epilepsy patients, and found that

Fig. 3 Physiological responses in the human fusiform gyrus show tolerance to strong occlusion. Invasive intracranial field potential recordings from an electrode in the fusiform gyrus in a subject with epilepsy (modified from Tang et al. 2014). This electrode responded selectively to faces (*left*, *gray* = individual trials, *green* = average of 9 repetitions). The other panels show single trial responses to five partial images of the face. *Black bar* indicates stimulus presentation time (150 ms). Despite heavy occlusion (89 %), the neural responses were similar to those obtained when presenting the whole object. However, the responses to partial images were significantly delayed. The *arrow* indicates latency of the response peak with respect to image onset

Fig. 4 Physiological responses to occluded objects were delayed compared to whole objects. Single trial decoding performance from a pseudopopulation of n = 60 electrodes for categorization (*black*, n = 5 categories) or identification (*gray*, n = 25 exemplars) (modified from (Tang et al. 2014)). The latency of selective information was delayed for occluded objects (*dashed line*) compared to whole objects (*solid line*)

317   neurons were surprisingly sensitive to even small degrees of occlusion (Rutishauser
318   et al. 2011). Their firing rates varied non-linearly with the amount of occlusion and
319   the responses to image parts did not necessarily bear a resemblance to the responses
320   to the whole images. These non-visual medial temporal lobe neurons, in effect,
321   'lost' the robustness developed in the ventral visual stream, in that individual parts
322   were not sufficient to drive the response to the level observed for whole faces.
323   In addition to the type of spatial integration demonstrated in the studies above,
324   the visual system is able to integrate information over time. Temporal integration is
325   particularly prominent in examples of action recognition. Yet, in some cases, dif-
326   ferent parts of the same object may appear in a dynamic fashion over time. At the

327 behavioral level, presenting object parts asynchronously significantly disrupts
328 object recognition performance, even when the temporal lag is as short as 16 ms
329 (Singer and Kreiman 2014). This disruption by asynchronous presentation is also
330 evident at the physiological level (Singer et al. 2015). Thus, the ability to spatially
331 integrate parts into a whole for recognition is quite sensitive to deviations from the
332 synchronous presentation of those parts.

333   In sum, early visual areas show evidence of contour completion in the presence
334 of both occluded and illusory contours when the corresponding edges are in close
335 spatial proximity. In some, but not all cases, these contour completion responses
336 show a delay with respect to both responses to real contours and responses in higher
337 visual areas. In higher visual areas responsible for object recognition, physiological
338 signals show strong robustness to large degrees of occlusion, consistent with
339 behavioral recognition performance, and these physiological signals also show a
340 significant delay. The robustness in the physiological responses and the dynamic
341 delays are consistent with the behavioral observations. These delays are interpreted
342 as originating from the involvement of recurrent and/or top-down connections
343 during the process of object completion. Given these initial steps in understanding
344 plausible neural circuits underlying recognition of occluded objects, we turn our
345 attention to describing the possible biological algorithms instantiated by these
346 neural signals.

347 # 4   Computational Models of Occluded Object Recognition

348 There has been significant progress over the last decade in developing computa-
349 tional models of object recognition (Deco and Rolls 2004; DiCarlo et al. 2012;
350 Kreiman 2013; Riesenhuber and Poggio 1999; Serre et al. 2007a). To a first
351 approximation, these models propose a hierarchical sequence of linear filtering and
352 non-linear max operations inspired by the basic principles giving rise to simple and
353 complex cells in primary visual cortex (Hubel and Wiesel 1962). Concatenating
354 multiple such operations gave rise to some of the initial models for object recog-
355 nition (Fukushima 1980). Recently, these ideas have also seen wide adoption in the
356 computer science literature in the form of deep convolutional neural networks (e.g.
357 (Krizhevsky et al. 2012) among many others). Both biologically inspired models
358 and deep convolutional neural networks (CNNs) geared towards performance on
359 benchmark datasets share similar core architectures.

4.1 Performance of Feed-Forward Models in Recognizing Occluded Objects

The canonical steps in feed-forward computational models is inspired by the observation of simple and complex cells in primary visual cortex of anesthetized cats. In their classic study, Hubel and Wiesel discovered 'simple' cells tuned to bars oriented at a particular orientation (Hubel and Wiesel 1959). They also described `complex' cells, which were also tuned to a preferred orientation, but exhibited a degree of tolerance to the spatial translation of the stimulus. They hypothesized that to generate this spatial invariance, the complex cells pool over simple cells whose receptive files tile the visual space with a max-like operation. This complex cell would then respond to an oriented bar regardless of its spatial location. Both hierarchical models of biological vision such as HMAX (Riesenhuber and Poggio 1999; Serre et al. 2007b) and CNNs are composed of alternating layers of tuning and pooling with increasingly more complicated tuning functions as one ascends this hierarchy. Whereas biologically-inspired models such as HMAX have about 4 layers, state-of-the-art computer vision models have moved to complex topologies with up to 20 layers and different mixtures of tuning and pooling layers (e.g. Russakovsky et al. 2015). Performance of feed-forward models such as HMAX on object recognition datasets match the pattern of human performance (Serre et al. 2007b). Additionally, the activity of individual layers in deep learning networks can capture the variance of neurons in the corresponding layers in macaque cortex (Cadieu et al. 2014; Yamins et al. 2014).

While these feed-forward architectures are designed to build tolerance to image transformations such as position and scale changes, they are not necessarily robust to the removal or occlusion of object features. Indeed, Fig. 5 shows the performance of an HMAX-like architecture in recognizing the same occluded objects form Figs. 2, 3 and 4. Small amounts of occlusion do not impair performance. However, performance drops rapidly with increasing occlusion, much more rapidly than human performance (see "behavior" line in Fig. 5 and compare Fig. 2 versus Fig. 5. Experimental simulations with other similar models confirm that that both HMAX and CNN models are challenged by recognition of occluded objects (Pepik et al. 2015; Wyatte et al. 2012b). Unlike position or scale transformations, the underlying representation in these models is not robust to occlusion. Feed-forward networks do not have explicit mechanisms to compensate for the missing features of occluded objects. In addition, because these models do not distinguish between the occluder and the object, the occluder can introduce spurious features that, through spatial pooling, are mixed with the object features. With small amounts of occlusion, the remaining features may be sufficient to lead to successful classification. However, with increasing levels of occlusion, the lack of sufficient information and completion mechanisms lead to a significant impairment in performance.
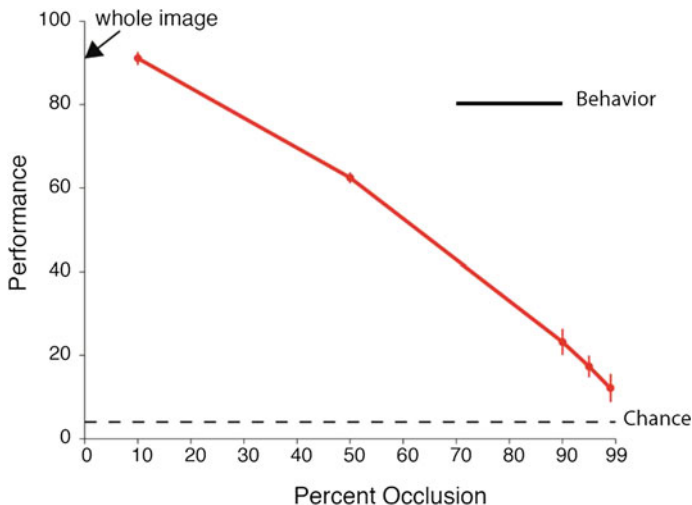
AQ2

Fig. 5 Challenge to feed-forward computational models Performance of a hierarchical feed-forward model of biological vision (Serre et al. 2007) on recognition of partial images similar to the one shown in Fig. 1f. *Dashed line* indicates chance level. Performance (*red line*) of the model is well below that of human subjects (*solid black line*) for heavy occlusion. Span of the *black line* indicates the range of difficulty tested for humans

4.2 Beyond Feed-Forward Models

Models that incorporate additional computations beyond feed-forward architectures can be subdivided into several categories. A group of computational models describes the process by which contours are amodally completed (Ullman 1976; Yuille and Kersten 2006). These models typically rely on an axiomatic set of desirable qualities that completed curves must satisfy (such as minimum total bending energy).

Another set of ideas argues that creating an understanding of the surfaces in a scene is critical to object recognition for occluded objects. Nakayama (1995) proposes a theory where surface representation is constructed via feedback in early visual cortex by learning associations between a viewed image and the underlying surface representations. The authors argue that this intermediate surface represen-tation is vital for subsequent recognition for occluded objects and mediates many important functions in texture segregation and visual search.

A different model proposes that neural representations of surfaces are created in three stages based on the low-level features (Sajda and Finkel 1995). First, points belonging to the same contour are bound together, followed by a process that determines the surfaces, and finally the surfaces are ordered by depth. Both feed-forward and feedback connections subserve to communicate between these three stages.

Several theories based on the role of feedback connections emphasize inference, but these ideas have largely not been operationalized into object recognition models. Predictive coding models are generative models of object recognition (Rao and Ballard 1999). In these models, higher visual areas send their predictions to lower levels, which then return only the mismatch between the predicted activity and the actual activity. This creates an efficient system where each layer only sends forwards signals that deviate from the receiving layer's predictions. The higher layers then attempt to generate the correct hypothesis of the image by reducing the incoming prediction errors. A related model proposes that visual cortex is essentially performing Bayesian inference where feed-forward inputs combine with top-down priors for recognition (Lee and Mumford 2003; Yuille and Kersten 2006).

As pointed out by (Wyatte et al. 2014), predictive coding models would expect occluded images to lead to increased activity in visual cortex, since the first generated hypotheses would have a very large prediction errors due to occlusion. Over time, we would expect this activity to subside as the system converged on an accurate hypothesis. Neurophysiological studies, however, find decreased or unchanged activity throughout visual cortex (Kovacs et al. 1995; Nielsen et al. 2006a; Tang et al. 2014).

Several models that more directly examine object recognition deal with recognition from partial information, and do not incorporate any of the amodal completion mechanisms previously described. These models take advantage of the extensive feedback and recurrent connections in visual cortex. While the role of these connections in attentional modulation has been extensively studied, their contribution to object recognition remains unclear. A particularly prominent and attractive class of models that can perform pattern completion is the all-to-all connectivity architectures such as Hopfield networks (Hopfield 1982). The Hopfield network generates attractors for previously learned patterns in such a way that if the network is initialized with partial information, it can dynamically evolve towards the right attractor. Interestingly, this type of dynamical convergence towards the attractor state could account for the type of delays observed in the behavioral and physiological experiments. This general principle is operationalized by a neural network model that combines bottom-up input with top-down signals carrying previously learned patterns to complete occluded objects (Fukushima 2005). For occluded patterns that are novel, this network attempts to interpolate from visible edges. The author applies this model to complete occluded letters of the alphabet. This concept has been extended to naturalistic objects with a feed-forward model that is augmented with recurrent feedback (Wyatte et al. 2012a). This recurrent feedback served to strengthen the feed-forward signals that were diminished from the occluded image.

Given that strong behavioral and neurophysiological evidence exists for amodal completion in human brain, and that surface representations are important for organizing the visual scene, theories of object recognition would be remiss to exclude these features in favor of purely feature-based recognition. An important step towards models that fully capture natural biological vision would be to integrate traditional feed-forward models with feedback mechanisms, including amodal

completion, surface generation, and top-down modulation based on priors and context. The challenge of recognizing occluded objects stands as the first test of these future integrative theories.

References

Allison T, Puce A, Spencer D, McCarthy G (1999) Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. Cereb Cortex 9:415–430

Bakin JS, Nakayama K, Gilbert CD (2000) Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. J Neurosci Off J Soc Neurosci 20:8188–8198

Bregman AS (1981) Asking the 'What for'question in auditory perception. In: Perceptual organization, pp 99–118

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10:e1003963

Callaway EM (2004) Feedforward, feedback and inhibitory connections in primate visual cortex. Neural Netw 17:625–632

Carpenter G, Grossberg S (2002) Adaptive resonance theory. In: The handbook of brain theory and neural networks. MIT Press, Cambridge

Chen J, Liu B, Chen B, Fang F (2009) Time course of amodal completion in face perception. Vis Res 49:752–758

Chen J, Zhou T, Yang H, Fang F (2010) Cortical dynamics underlying face completion in human visual system. J Neurosci Off J Soc Neurosci 30:16692–16698

Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. Curr Opin Neurobiol 17:140–147

Deco G, Rolls ET (2004) A neurodynamical cortical model of visual attention and invariant object recognition. Vis Res 44:621–642

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

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434

Doniger GM, Foxe JJ, Murray MM, Higgins BA, Snodgrass JG, Schroeder CE, Javitt DC (2000) Activation timecourse of ventral visual stream object-recognition areas: high density electrical mapping of perceptual closure processes. J Cogn Neurosci 12:615–621

Douglas RJ, Martin KA (2004) Neuronal circuits of the neocortex. Annu Rev Neurosci 27:419–451

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47

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

Fukushima K (2005) Restoring partly occluded patterns: a neural network model. Neural Netw 18:33–43

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vis Res 41:2261–2271

Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. Vis Res 41:1409–1422

Gross C, Bender D, Rocha-Miranda C (1969) Visual receptive fields of neurons in inferotemporal cortex of the monkey. Science 166:1303–1306

Hegde J, Fang F, Murray S, Kersten D (2008) Preferential responses to occluded objects in the human visual cortex. J Vis 8:1–16

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. PNAS 79:2554–2558

Hubel D, Wiesel T (1959) Receptive fields of single neurons in the cat's striate cortex. J Physiol (Lond) 148:574–591

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154

Hung C, Kreiman G, Poggio T, DiCarlo J (2005) Fast read-out of object identity from macaque inferior temporal cortex. Science 310:863–866

Issa EB, Dicarlo JJ (2012) Precedence of the eye region in neural processing of faces. J Neurosci Off J Soc Neurosci 32:16666–16682

Ito M, Tamura H, Fujita I, Tanaka K (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. J Neurophysiol 73:218–226

Johnson JS, Olshausen BA (2005) The recognition of partially visible natural objects in the presence and absence of their occluders. Vis Res 45:3262–3276

Kanizsa G (1979) Organization in vision: essays on gestalt perception. Praeger Publishers

Kellman PJ, Guttman SE, Wickens TD (2001) Geometric and neural models of object. In: From fragments to objects: segmentation and grouping in vision, vol 130, p 183

Keysers C, Xiao DK, Foldiak P, Perret DI (2001) The speed of sight. J Cogn Neurosci 13:90–101

Komatsu H (2006) The neural mechanisms of perceptual filling-in. Nat Rev Neurosci 7:220–231

Kosai Y, El-Shamayleh Y, Fyall AM, Pasupathy A (2014) The role of visual area V4 in the discrimination of partially occluded shapes. J Neurosci Off J Soc Neurosci 34:8570–8584

Kovács G, Vogels R, Orban GA (1995) Selectivity of macaque inferior temporal neurons for partially occluded shapes. J Neurosci Off J Soc Neurosci 15:1984–1997

Kreiman G (2013) Computational models of visual object recognition. In: Panzeri S, Quian Quiroga R (eds) Principles of neural coding. Taylor and Fracis Group

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20:1434–1448

Lee TS, Nguyen M (2001) Dynamics of subjective contour formation in the early visual cortex. Proc Natl Acad Sci USA 98:1907–1911

Lerner Y, Harel M, Malach R (2004) Rapid completion effects in human high-order visual areas. Neuroimage 21:516–526

Lerner Y, Hendler T, Malach R (2002) Object-completion effects in the human lateral occipital complex. Cereb Cortex 12:163–177

Liu H, Agam Y, Madsen JR, Kreiman G (2009) Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. Neuron 62:281–290

Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5:552–563

Logothetis NK, Sheinberg DL (1996) Visual object recognition. Annu Rev Neurosci 19:577–621

Marr D (1982) Vision. Freeman Publishers, San Francisco

Mel B (1997) SEEMORE: combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. Neural Comput 9:777

Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern 66:241–251

Murray MM (2004) Setting boundaries: brain dynamics of modal and amodal illusory shape completion in humans. J Neurosci 24:6898–6903

Murray RF, Sekuler AB, Bennett PJ (2001) Time course of amodal completion revealed by a shape discrimination task. Psychon Bull Rev 8:713–720

Nakayama K, He Z, Shimojo S (1995) Visual surface representation: a critical link between lower-level and higher-level vision. In: Kosslyn S, Osherson D (eds) Visual cognition. The MIT Press, Cambridge

Nielsen K, Logothetis N, Rainer G (2006a) Dissociation between LFP and spiking activity in macaque inferior temporal cortex reveals diagnostic parts-based encoding of complex objects. J Neurosci 26:9639–9645

Nielsen KJ, Logothetis NK, Rainer G (2006b) Discrimination strategies of humans and rhesus monkeys for complex visual displays. Curr Biol CB 16:814–820

Olshausen BA, Anderson CH, Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J Neurosci Off J Soc Neurosci 13:4700–4719

Olson IR, Gatenby JC, Leung HC, Skudlarski P, Gore JC (2004) Neuronal representation of occluded objects in the human brain. Neuropsychologia 42:95–104

Pepik B, Benenson R, Ritschel T, Schiele B (2015) What is holding back convnets for detection? arXiv:150802844

Perrett D, Hietanen J, Oeam M, Benson P (1992) Organization and functions of cells responsive to faces in the temporal cortex. Phil Trans Roy Soc 355:23–30

Peterhans E, von der Heydt R (1991) Subjective contours - bridging the gap between psychophysics and physiology. Trends Neurosci 14:112–119

Potter M, Levy E (1969) Recognition memory for a rapid sequence of pictures. J Exp Psychol 81:10–15

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87

Rauschenberger R, Peterson MA, Mosca F, Bruno N (2004) Amodal completion in visual search: preemption or context effects? Psychol Sci 15:351–355

Richmond B, Wurtz R, Sato T (1983) Visual responses in inferior temporal neurons in awake Rhesus monkey. J Neurophysiol 50:1415–1432

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025

Rolls E (1991) Neural organization of higher visual functions. Curr Opin Neurobiol 1:274–278

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015). Imagenet large scale visual recognition challenge. Int J Comput Vis

Rutishauser U, Tudusciuc O, Neumann D, Mamelak AN, Heller AC, Ross IB, Philpott L, Sutherling WW, Adolphs R (2011) Single-unit responses selective for whole faces in the human amygdala. Curr Biol CB 21:1654–1660

Sajda P, Finkel LH (1995) Intermediate-level visual representations and the construction of surface perception. J Cogn Neurosci 7:267–291

Schmolesky M, Wang Y, Hanes D, Thompson K, Leutgeb S, Schall J, Leventhal A (1998) Signal timing across the macaque visual system. J Neurophysiol 79:3272–3278

Sehatpour P, Molholm S, Schwartz TH, Mahoney JR, Mehta AD, Javitt DC, Stanton PK, Foxe JJ (2008) A human intracranial study of long-range oscillatory coherence across a frontal-occipital-hippocampal brain network during visual object processing. Proc Natl Acad Sci USA 105:4399–4404

Sekuler AB, Murray RF (2001) Amodal completion: a case study in grouping. Advances in Psychology 130:265–293

Sekuler AB, Palmer SE (1992) Perception of partly occluded objects: a microgenetic analysis. J Exp Psychol Gen 121:95–111

Sekuler AB, Palmer SE, Flynn C (1994) Local and global processes in visual completion. Psychol Sci 5:260–267

Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T (2007a) A quantitative theory of immediate visual recognition. Prog Brain Res 165C:33–56

Serre T, Oliva A, Poggio T (2007b) Feedforward theories of visual cortex account for human performance in rapid categorization. PNAS 104:6424–6429

Shimojo S, Nakayama K (1990a) Amodal representation of occluded surfaces: role of invisible stimuli in apparent motion correspondence. Perception 19:285–299

Shimojo S, Nakayama K (1990b) Real world occlusion constraints and binocular rivalry. Vis Res 30:69–80

Shore DI, Enns JT (1997) Shape completion time depends on the size of the occluded region. J Exp Psychol Hum Percept Perform 23:980–998

Singer JM, Kreiman G (2014) Short temporal asynchrony disrupts visual object recognition. J Vis 14:7

Singer JM, Madsen JR, Anderson WS, Kreiman G (2015) Sensitivity to timing and order in human visual cortex. J Neurophysiol 113:1656–1669

Singh M (2004) Modal and amodal completion generate different shapes. Psychol Sci 15:454–459

Sugita Y (1999) Grouping of image fragments in primary visual cortex. Nature 401:269–272

Sun Y, Wang X, Tang X (2014) Deeply learned face representations are sparse, selective, and robust. arXiv:14121265

Taigman Y, Yang M, Ranzato MA, 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), pp. 1701–1708. IEEE

Tanaka K (1996) Inferotemporal cortex and object vision. Annu Rev Neurosci 19:109–139

Tang H, Buia C, Madhavan R, Crone NE, Madsen JR, Anderson WS, Kreiman G (2014) Spatiotemporal dynamics underlying object completion in human ventral visual cortex. Neuron 83:736–748

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520–522

Ullman S (1976) Filling-in the gaps: the shape of subjective contours and a model for their generation. Biol Cybern 25:1–6

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

Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. Prog Neurobiol 51:167–194

Wyatte D, Curran T, O'Reilly R (2012a) The limits of feedforward vision: recurrent processing promotes robust object recognition when objects are degraded. J Cogn Neurosci 24:2248–2261

Wyatte D, Jilk DJ, O'Reilly RC (2014) Early recurrent feedback facilitates visual object recognition under challenging conditions. Front Psychol 5:674

Wyatte D, Tang H, Buia C, Madsen J, O'Reilly R, Kreiman G (2012b) Object completion along the ventral visual stream: neural signatures and computational mechanisms. In: Computation and systems neuroscience, Salt Lake City, Utah

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci USA 111:8619–8624

Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10:301–308

Zipser K, Lamme VA, Schiller PH (1996) Contextual modulation in primary visual cortex. J Neurosci 16:7376–7389

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	Please confirm if the section headings identified are correct.	
AQ2	References 'Nielsen et al. (2006), Serre et al. (2007)' are cited in the text but not provided in the reference list. Please provide the respective references in the list or delete these citations.	
AQ3	As Refs. [35] of Kovács (1995) and [36] of Kovács (1995) are same, we have deleted the duplicate references and renumbered accordingly. Please check and confirm.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	・・・ under matter to remain	Ⓙ
Insert in text the matter indicated in the margin	⋏	New matter followed by ⋏ or ⋏⊘
Delete	/ through single character, rule or underline or ⊢——⊣ through all characters to be deleted	⌀ or ⌀⊘
Substitute character or substitute part of one or more word(s)	/ through letter or ⊢——⊣ through characters	new character / or new characters /
Change to italics	— under matter to be changed	‿
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	= under matter to be changed	=
Change to bold type	∿ under matter to be changed	∿
Change to bold italic	≈ under matter to be changed	≋
Change to lower case	Encircle matter to be changed	≢
Change italic to upright type	(As above)	⊥
Change bold to non-bold type	(As above)	⊥
Insert 'superior' character	/ through character or ⋏ where required	⋎ or ⋏ under character e.g. ⋎² or ⋏²
Insert 'inferior' character	(As above)	⋏ over character e.g. ⋏₂
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	⋎ or ⋏ and/or ⋎ or ⋏
Insert double quotation marks	(As above)	⋎ or ⋏ and/or ⋎ or ⋏
Insert hyphen	(As above)	⊢⊣
Start new paragraph	⌐	⌐
No new paragraph	⌒	⌒
Transpose	⊔⊓	⊔⊓
Close up	linking ⌒ characters	⌒
Insert or substitute space between characters or words	/ through character or ⋏ where required	Ⱶ
Reduce space between characters or words	| between characters or words affected	↑