Cognition 170 (2018) 228–244

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Attention to distinguishing features in object recognition: An interactive-iterative framework

Orit Baruch[*], Ruth Kimchi, Morris Goldsmith

University of Haifa, Israel

ARTICLE INFO

Keywords:
Object recognition
Object categorization
Visual attention
Distinguishing features
Interactive
Iterative
Top-down processes
Bottom-up processes

ABSTRACT

This article advances a framework that casts object recognition as a process of discrimination between alternative object identities, in which top-down and bottom-up processes interact—iteratively when necessary—with attention to distinguishing features playing a critical role. In two experiments, observers discriminated between different types of artificial fish. In parallel, a secondary, variable-SOA visual-probe detection task was used to examine the dynamics of visual attention. In Experiment 1, the fish varied in three distinguishing features: one indicating the general category (saltwater, freshwater), and one of the two other features indicating the specific type of fish within each category. As predicted, in the course of recognizing each fish, attention was allocated iteratively to the distinguishing features in an optimal manner: first to the general category feature, and then, based on its value, to the second feature that identified the specific fish. In Experiment 2, two types of fish could be discriminated on the basis of either of two distinguishing features, one more visually discriminable than the other. On some of the trials, one of the two alternative distinguishing features was occluded. As predicted, in the course of recognizing each fish, attention was directed initially to the more discriminable distinguishing feature, but when this feature was occluded, it was then redirected to the less discriminable feature. The implications of these findings, and the interactive-iterative framework they support, are discussed with regard to several fundamental issues having a long history in the literatures on object recognition, object categorization, and visual perception in general.

1. Introduction

Object recognition is of fundamental importance for the perception of and interaction with our environment. Despite extensive research, however, there is still no complete and comprehensive theory that can explain how we recognize objects, and some of the most basic characteristics of the recognition process continue to be a subject of debate.

One controversial issue concerns the role of top-down versus bottom-up processing. Despite many other differences, classic theories of object recognition (e.g., Biederman, 1987; Marr & Nishihara, 1978; Poggio & Edelman, 1990; Reisenhuber & Poggio, 1999; Tarr & Bülthoff, 1995; Tarr & Pinker, 1989; Ullman, 1989) are generally united in what might be called the orthodox view—that object recognition is based primarily on a bottom-up analysis of the visual input; recognition is achieved when some temporary representation of the input image matches a stored object representation. The functional architecture of the visual cortex—the increase in the receptive field size and in representational complexity from lower to higher areas in the cortex (Maunsell & Newsome, 1987; Vogels & Orban, 1996)—has been pointed to as consistent with the bottom-up view. Also, findings that the processes involved in object recognition are sometimes remarkably fast, occurring within 100–200 ms of stimulus presentation (e.g., Thorpe, Fize, & Marlot, 1996), have been taken by some researchers as evidence that object recognition can occur largely with feed-forward processing alone (e.g., Wallis & Rolls, 1997; but see Evans & Treisman, 2005).

Some proposals, however, have challenged the orthodox view, emphasizing the need for both bottom-up and top-down processing (e.g., Bar, 2003; Bullier, 2001; Ganis, Schendan, & Kosslyn, 2007; Humphreys, Riddoch, & Price, 1997; Lee, 2002; McClelland & Rumelhart, 1981; Schendan & Maher, 2009; Schendan & Stern, 2008; Ullman, 1995). For example, Bar (2003), Bar et al., 2006), inspired by Ullman's (1995) model, proposed that partially processed visual data based on low spatial frequencies of the input is transmitted from the initial areas of the visual stream directly to the orbito-frontal cortex. This low-spatial-frequency representation invokes initial hypotheses regarding the identity of the input, which subsequently facilitate the identification process by constraining the number of possibilities that have to be inspected (see also

[*] Corresponding author.
E-mail address: oritb@research.haifa.ac.il (O. Baruch).

http://dx.doi.org/10.1016/j.cognition.2017.10.007
Received 15 September 2016; Received in revised form 3 October 2017; Accepted 6 October 2017

Peyrin et al., 2010). Similarly, Bullier (2001) proposed a model of visual processing by which initially, information from the visual stimulus is transferred rapidly via magnocellular, dorsal pathways. Results from this first-pass computation are then sent back by feedback connections and used to guide further processing of parvocellular and koniocellular information in the inferotemporal cortex. The existence of massive projections from higher to lower areas of the visual pathways (e.g., Bullier, 2001; Lamme & Roelfsema, 2000) suggests that the involvement of top-down processing in object recognition is physiologically viable. Top-down influences on object recognition are also implicated in behavioral studies. For example, advance information about the target in RSVP experiments improves target detection (Intraub, 1981), priming by category names substantially improves object identification (Reinitz, Wright, & Loftus, 1989), and objects are recognized better in expected than in unexpected contexts (e.g., Bar & Ullman, 1996; Biederman, 1972, 1981).

Several models propose that top-down and bottom-up information might be integrated via an iterative error-minimization mechanism, where top-down predictions are matched to processed bottom-up information in recursive, interacting loops of activity (Friston, 2005; Hinton, Dayan, Frey, & Neal, 1995; Kveraga, Ghuman, & Bar, 2007; Mumford, 1992; Ullman, 1995).

2. Role of attention in object recognition

Partly related to the preceding issue is an ongoing controversy regarding the role of attention in object recognition. Some researchers have provided evidence suggesting that object recognition can be carried out in the near absence of attention (e.g., Li, VanRullen, Koch, & Perona, 2002; Luck, Vogel, & Shapiro, 1996). Other researchers, however, hold that attention plays a central role (e.g., Ganis & Kosslyn, 2007; Hochstein & Ahissar, 2002; Treisman & Gelade, 1980). Most notably, the highly influential Feature Integration Theory (Treisman & Gelade, 1980) holds that attention is crucial for the perception of an integrated object, as it operates to bind featural information represented in independent feature maps. In contrast, the more recent Reverse Hierarchy Theory (Hochstein & Ahissar, 2002) holds that whereas the initial perception of coherent conjoined objects can be achieved "at a glance" under spread attention, based on feed-forward processing alone, top-down focused attention must subsequently be invoked to consciously identify specific details such as orientation, color, and precise location.

An additional role for attention in object recognition has emerged from the view of visual perception as a process of hypothesis testing (Gregory, 1966; von Helmholtz, 1867), by which attention is directed to diagnostic feature information that is used to decide between alternative hypotheses regarding object identity (e.g., Baruch, Kimchi, & Goldsmith, 2014; Ganis & Kosslyn, 2007; Ganis et al., 2007). This view, advanced in the present research, is outlined in the following section.

3. Interactive-Iterative attentional framework for object recognition

The present work was guided by a framework that views object recognition as a process of discrimination between probable alternatives—a process in which bottom-up and top-down processes interact, iteratively when necessary, with attention playing a crucial role in this interaction. We outline here the set of principles that comprises this framework (a more concrete schematic depiction appears as Fig. 13 in General Discussion)—essentially, a synthesis of ideas that have been proposed previously, from which specific predictions can be derived and empirically examined.

3.1. Object recognition begins with expectations based on past experience and present context

Object recognition undoubtedly requires an analysis of visual data. Yet, contrary to the conventional view, we suggest that the recognition process actually begins at the top. Everyday situations generally evoke expectations about probable objects, based on world knowledge, context, and goals (e.g., Bar, 2004; Biederman, 1972; Norman & Bobrow, 1976; Palmer, 1975). Even in the laboratory, expectations are evoked by the experimental task. Pure data-driven recognition—where an object could be anything—are presumably quite rare, and can be seen as a special case in which the probable alternatives are all objects known to the observer. A similar view of perception has recently been revived in several models using Bayesian inference, in which top-down priors help to disambiguate noisy bottom-up sensory input signals (e.g., Epshtein, Lifshitz, & Ullman, 2008; Friston & Kiebel, 2009).

3.2. The initial visual input is inherently limited

The initial information extracted from the visual scene in a data-driven (bottom-up) manner is inherently partial. In natural scenes, portions of objects—those on the side away from the viewer—are hidden from view and surfaces may undergo occlusion; sometimes the viewing conditions are poor, and at other times the relevant diagnostic information is subtle and cannot be acquired at a glance. Moreover, even under optimal viewing conditions, the initial information may be partial (e.g., coarse information carried by low spatial frequencies; Bar, 2003; Fabre-Thorpe, 2011; Hughes, Nozawa, & Kitterle, 1996). Although, depending on context, the initial partial information may sometimes suffice for recognition, in many cases object recognition will require additional processing.

3.3. Perceptual hypotheses guide the allocation of attention to distinguishing features

It was suggested long ago (Gregory, 1966; von Helmholtz, 1867) that perception is essentially a hypothesis-assessment process. Building on this idea, and in line with more recent ideas concerning the "predictive brain" (e.g., Bubic, von Cramon, & Schubotz, 2010; Enns & Llreas, 2008), we assume that the observer's expectations—-whether formed prior to or in interaction with the visual input—evoke a set of alternative hypotheses regarding the possible identity of the observed object. These hypotheses are expressed as the activation of internal representations of candidate objects, giving special weight to diagnostic features[1] (e.g., Gillebert, Op de Beeck, Panis, & Wagemans, 2009; Schyns & Rodet, 1997; Sigala & Logothetis, 2002; Wagar & Dixon, 2005) that discriminate between competing hypotheses. Attention is then directed to these distinguishing features in order to facilitate the extraction of the relevant information (see also Ganis & Kosslyn, 2007; Kosslyn, 1994).

The specific claim that attention is directed to distinguishing features in object recognition has been empirically addressed in relatively few studies, most of which used eye tracking as an indirect measure of spatial attention. For example, Rehder and Hoffman (2005a, 2005b; see also Blair, Watson, Walshe, and Maj, 2009) found that during visual object category learning, diagnostic features were fixated more often than non-diagnostic features, and that the proportion of correct responses correlated with the time diagnostic features were fixated. Using a more direct measure of spatial attention in the context of word and

[1] Note that the notion of features (and hence, distinguishing features) as conceived here is very broad, and refers to any aspect of an object that can serve to discriminate between the set of probable alternatives. Such aspects may include, for example, structural or configural features (e.g., *geons*; Biederman, 1987), surface features (e.g., color or texture), global features (e.g. global shape: elongated vs. round), or localized features and parts (e.g., the shape or color of a beak).

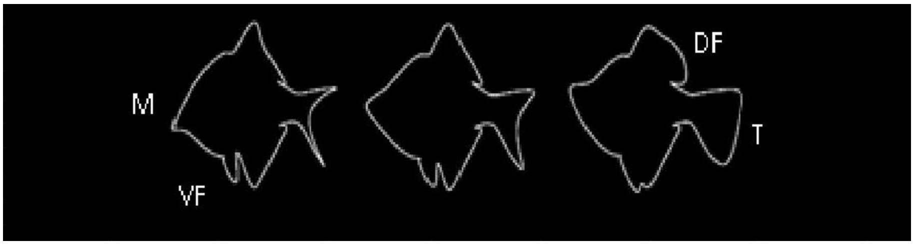

Fig. 1. Examples of fish stimuli used in the experiments. The fish have four basic feature variables: Mouth (M), Tail (T), Dorsal Fin (DF) and Ventral Fin (VF), with three values each.

letter recognition, Navon and Margalit (1983) found that detection rate of a visual probe was highest when the probe appeared near the feature that distinguished between two competing word or letter alternatives.

Recently, utilizing several different methods, including primed matching, visual probe detection, and spatial cueing, Baruch et al. (2014) provided further converging evidence for the claim that attention is allocated to distinguishing features in the course of object recognition. Using various sets of line drawings of artificial fish, they showed that: (1) recognizing a fish primed its distinguishing features but not its other features, (2) visual probes presented near a distinguishing feature were detected faster and more often than probes presented near a non-distinguishing feature, and (3) advance allocation of attention to the location of the distinguishing feature by a transient precue yielded faster recognition latencies than when the location of a non-distinguishing feature was cued. Furthermore, they showed that the attended distinguishing features are context dependent: Attention was allocated to different features of the same fish, depending on the overall stimulus set. Similar results were found using photographs of natural sea animals, despite substantial variation in physical characteristics, posture, lighting, viewing angle, and so forth.

3.4. Object recognition is an interactive iterative process

Information extracted from the initially attended features may suffice for recognition, but if not, hypotheses are refined and the process is repeated in an iterative manner until recognition is achieved. Although there may be instances in which the visual data quality is high and the set of possible alternatives highly constrained so that recognition is seemingly instantaneous (e.g., Thorpe et al., 1996), achieved in a single iteration of bottom-up analysis, by our view such instances are best conceived as a special case of what is potentially a more complex (interactive and iterative) object recognition process (see General Discussion).

Although the idea that the object recognition process may involve interactive-iterative shifts of attention to diagnostic visual information has been put forward in theoretical discussions (e.g., Ganis & Kosslyn, 2007), surprisingly, almost no empirical evidence for this idea has been provided. To our knowledge, there is only one study, using eye movement measures, whose results can be taken to suggest that attention is allocated to distinguishing features in an iterative manner. In an artificial-object categorization task, Blair et al. (2009) found systematic temporal patterns in the shifting of eye fixations, suggesting that the information gleaned from fixating on one distinguishing feature could be used to determine which other distinguishing feature would be most informative of category identity, and therefore fixated next. These results are important and suggestive, but as stated by the authors themselves, because they are based on overt eye movements, they are limited in their implications regarding the role and dynamics of covert visual attention.

4. The present study

In this article we report two experiments designed to examine the idea that object recognition is an interactive iterative process, in which hypotheses about probable object identity drive the allocation of attention to the relevant distinguishing features, and if necessary, redirect attention to additional distinguishing features in an iterative manner until a specific hypothesis is confirmed. This general proposition was examined in two different object recognition (categorization) tasks,[2] in which the initially extracted information is limited, either because the relevant distinguishing features are subtle and spatially disparate (and therefore cannot be extracted all at once; Experiment 1), or because the most discriminable distinguishing feature is sometimes occluded (Experiment 2). We used a secondary visual-probe detection task (Baruch et al., 2014), with probes presented at several different SOAs, to examine the dynamics of visual attention in each of these situations.

4.1. General method

In both experiments, stimulus presentation and data collection were driven by a computer workstation with 17″, 1024 × 768 resolution monitor.

4.1.1. Recognition task

The object stimuli were two-dimensional line drawings of fish, adopted from Sigala and Logothetis (2002). These artificial objects allow maximal experimental control over the object features, and hence enable precise predictions regarding the dynamics of attention in the object recognition task, yet they represent natural objects. The fish have four local shape features: Mouth (M), Tail (T), Dorsal Fin (DF) and Ventral Fin (VF), with three values each (Fig. 1). They also have additional features such as texture and color. In each experiment, a subset of the features was used to define (implicitly) different types of fish, whose names were chosen to sound like real fish names (e.g., Grout or Tass).[3] On each trial, a specific instance of a fish, belonging equally often to one of the relevant fish types, was presented at the center of the screen. Participants were required to recognize the type of fish and instructed to respond with the appropriate key press as rapidly and accurately as possible. The fish remained on the screen until the participant responded. A short low-frequency audio tone was provided as feedback on incorrect response trials.

4.1.2. Secondary probe-detection task

Our basic research strategy was to use the pattern of allocations of attention in a controlled situation as a window into the dynamics of the processes that underlie object recognition. To examine this pattern, in

[2] In line with the general theoretical framework promoted in this article, object recognition and object categorization tasks are treated as essentially equivalent—both involve identifying a presently viewed visual stimulus as an instance of a particular object type or category, from among a set of expected-probable object types or categories. This point will be addressed further at the beginning of the General Discussion section.

[3] Although highly advantageous in terms of experimental control, the use of object line drawings with well-defined visual features may raise concerns regarding the generalizability to more naturalistic recognition situations, which are inherently more "noisy" both in the overall visual data and in the distinguishing features themselves. To partly address this issue, in a previous study (Baruch et al., 2014,) we included an experiment using photographs of real sea mammals (seals, sea lions and sea elephants). The results of that experiment showed that the allocation of attention to distinguishing features also holds for the categorization of pictures of natural objects despite substantial within-category variation in physical characteristics, posture, viewing angle, and more. In addition, post-session interviews revealed that in distinguishing seals from sea lions, none of the participants were able to explicitly verbalize the distinguishing feature to which they were actually attending.

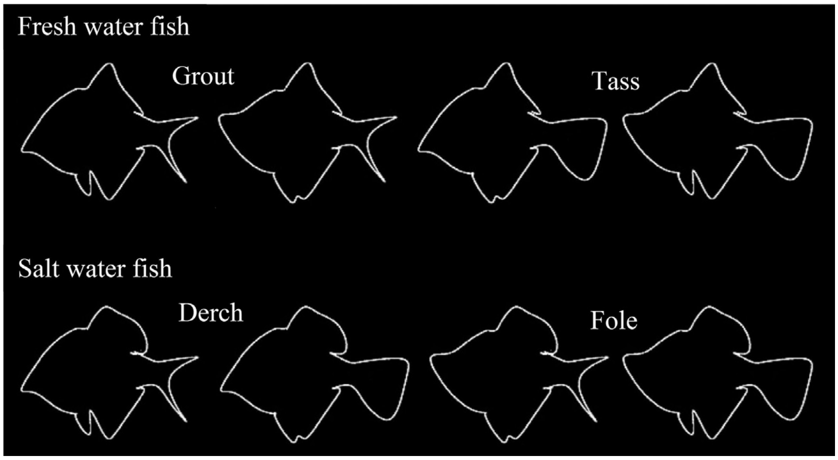

Fig. 2. Example of the outlines of fish stimuli that were used in Experiment 1. Freshwater fish (concave DF) are distinguished from saltwater fish (convex DF) by their DF. Freshwater fish ("Grout" and "Tass") are distinguished by their tail (T pointed or rounded); Saltwater fish ("Fole" and "Derch") are distinguished by their mouth (M upturned or downturned).

parallel to the primary object recognition task, a secondary probe detection task was used, combined with a manipulation of probe SOA. In addition to the fish object that was displayed until the participant's recognition response, on half of the trials a visual probe was displayed briefly next to one of the fish features at one of several different SOAs. Participants were instructed that their primary task was to identify the presented type of fish by pressing the appropriate key. If they noticed a probe, however, they were instructed to indicate the presence of the probe by pressing the recognition key twice ("double press") rather than once. This double-press response method (taken from Baruch et al., 2014) is designed to reduce any potential conflict that might arise when providing two separate responses at once: There is no confusion regarding the order of responses or which keys to press—the selected recognition key is either pressed once, indicating the recognition response alone, or double-pressed (similar to double-clicking a mouse), indicating both the recognition decision and that a probe was detected. The probes themselves, which differed between the two experiments, were designed such that (a) their onset and offset would not capture attention, and (b) their detection would be very difficult when they appeared outside the focus of attention. Thus, observed changes in probe detection rate at particular feature locations at particular SOAs could be used to track changes in the allocation of spatial attention between the different fish features during the course of the recognition process.

4.1.3. Training phase

Both experiments included a training phase at the beginning of the experiment. The training phase was used to familiarize the participant with the types of fish relevant to the experiment, the corresponding response keys, and the secondary probe-detection task, including single- versus double key presses. In order to facilitate learning of the dual recognition and probe-detection tasks, probes appeared on 80% of the trials in the training phase. In both experiments, this phase ended when the participant had completed at least 60 trials and made no more than two recognition errors on 20 consecutive trials (i.e., 90% accuracy).

In both experiments, the training phase was followed by an additional block of practice trials that were identical to the experimental trials in all respects.

4.1.4. Participants

The participants in both experiments were undergraduate students, all with normal, or corrected to normal vision. The chosen sample sizes were found to be sufficient to reveal the relevant effects in a previous study (Baruch et al., 2014) using similar methods in a completely within-participants design. Different participants were included in each experiment.

5. Experiment 1

This experiment examined the dynamics of the allocation of attention in an object recognition task in which the visual information that could be derived in a single glance would be insufficient to conclusively identify the object. This situation was created by defining four different types of fish that varied in three distinguishing features: one indicating the general family (salt-water, fresh-water), and one of the two other features indicating the specific type of fish within each category (see Fig. 2). The three distinguishing features were the shapes of the dorsal fin (DF), mouth (M), and tail (T). A fourth feature, the shape of the ventral fin (VF), was unrelated to object identity. Because the three local distinguishing features were subtle and spatially disparate, we expected that attention would need to be focused sequentially in order to extract all of the relevant visual information. Furthermore, by organizing the four specific types of fish using a nested family structure (see also Blair et al., 2009, Experiment 2), we created a situation in which there was an optimal sequence of attentional allocations that would minimize the number of distinguishing features that need to be examined in order to conclusively identify each fish: There was one "pivot" feature, the DF, whose value determined the general family (freshwater or saltwater) the fish belonged to, and accordingly, which of the other two potential distinguishing features (mouth or tail) was in fact diagnostic of the identity of the presented fish. Therefore, assuming that only one feature can be attended at a time, the most efficient algorithm would be to direct attention initially to the DF, and then, according to its value (shape), shift attention to the second relevant distinguishing feature (mouth or tail) whose value would conclusively identify the fish (see Fig. 3). It is assumed that in general, the visual system learns the most efficient "algorithm" for recognizing an object, and that the same "algorithm" is used each time the object is observed. The latter assumption is in line with results indicating that the pattern of saccadic eye movements tends to be repeated by a specific observer when inspecting a given object (Yarbus, 1967)

Note that in the present case, because the subsequent shift of attention to a particular distinguishing feature depends on the value of the initially attended feature, the optimal sequence of attentional allocations involves an iterative interaction between bottom-up and top-down processes. Alternatively, however, the fish recognition task could also be performed by testing the distinguishing features in a random or arbitrary sequence, in which case two or three attentional fixations would be needed, depending on the order in which the features are examined and the particular type of fish that is presented. Finally, if we are wrong in our assumption that focused attention is needed to extract the relevant information at each feature location, two or perhaps all three distinguishing features might be examined simultaneously, under spatially spread attention.

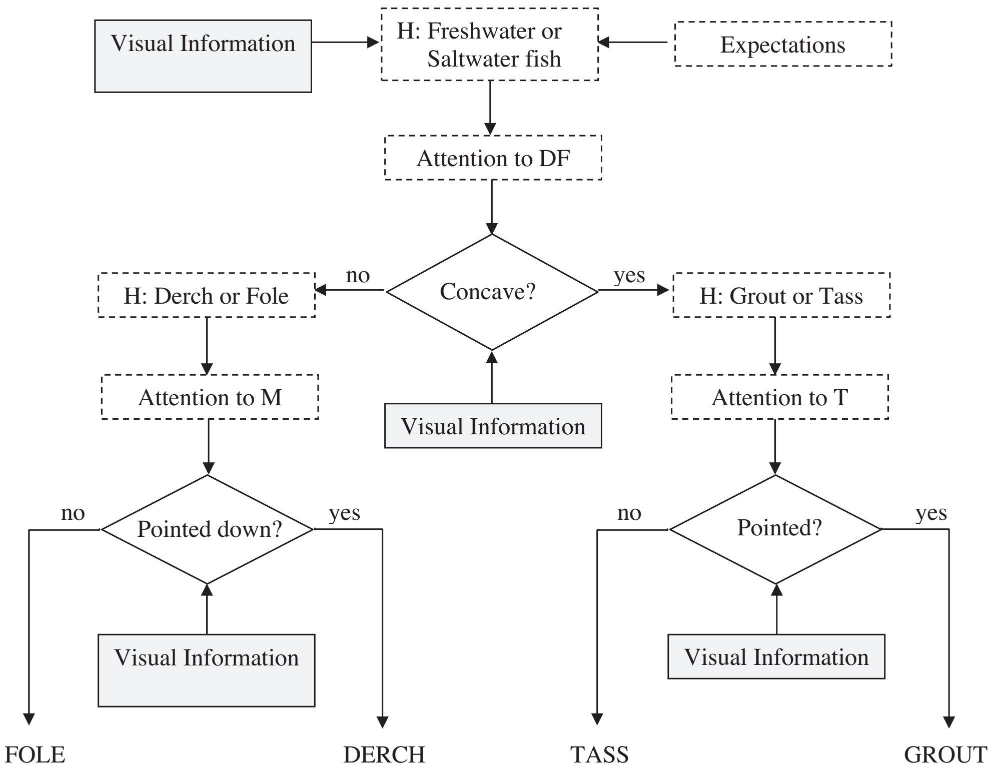

Fig. 3. Schematic diagram of the predicted interactive-iterative fish recognition process in Experiment 1. The initial hypothesis (H: the fish is either saltwater or freshwater) initiates top-down direction of attention to the most informative feature, DF. Visual information extracted from the attended DF feature evokes a more refined hypothesis regarding a specific fish category, leading to top-down direction of attention to the relevant additional diagnostic feature (M or T). The visual information extracted by attending to this feature is sufficient to recognize the specific type of fish. Note that given the nested hierarchical organization of the stimulus set, initial attention to the DF "pivot" feature ensures that only two attentional fixations are required to achieve recognition.

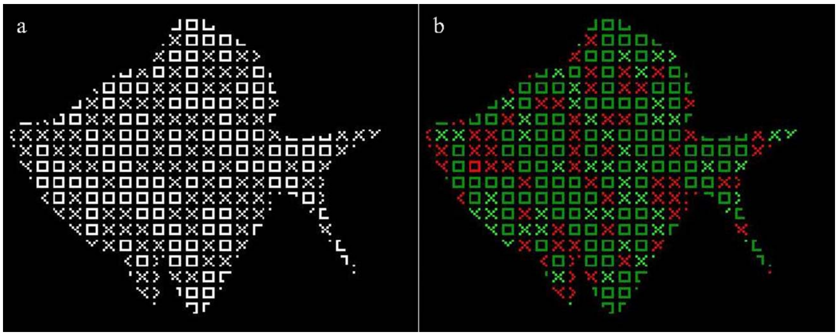

Fig. 4. Examples of gray and corresponding colored textured fish stimuli used in Experiment 1. The fish were initially displayed in gray (panel a), after which color was added for 400 ms (panel b), after which the display returned to gray. The red square element located near the mouth in panel b, is an example of the color-shape conjunction probe used in the secondary probe-detection task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As described earlier, a secondary variable-SOA probe detection task, performed together with the primary fish recognition task, was used to assess these different possibilities. For half of the presented fish, the probe appeared at one of the four feature locations (DF, VF, M, T) at one of four fish-probe SOAs, ranging from 0 ms to 450 ms. In this experiment, the probe was a single red square appearing among a large number of green squares and red and green Xs that comprised the "texture" of each fish (to prevent capture of attention by probe onset/offset, the red or green color was added to all texture elements simultaneously, which were otherwise colored gray; see Fig. 4). Essentially, then, the probe was a color-shape conjunction target whose detection generally requires spatially focused attention (Treisman & Gelade, 1980). Hence, we assumed that the probe would be detected only, or primarily, when the dynamics of attentional allocations in the course of recognizing the fish caused the focus of attention to coincide with the location of the probe, at the time the probe was presented.

If, during the course of recognizing each fish, attention is allocated to the distinguishing features sequentially according to the optimal (interactive-iterative) algorithm, an interaction between probe location and SOA is expected, such that at shorter SOAs probe detection will be best at the location of the DF (the initially attended, pivot feature), whereas at longer SOAs it will be best at the location of the other relevant distinguishing feature (M or T, depending on the family of the fish). Put another way, the point in time at which the detection rate for probes at the DF is maximal should correspond to the initial determination of fish family (freshwater/saltwater), whereas the maximal detection rate for probes at the second relevant distinguishing feature (M or T) should correspond to the subsequent determination of the specific type of fish. Probe detection rate at the VF (which is irrelevant for recognizing the fish) is expected be the worst overall, and not to vary with SOA (see Fig. 5).

Alternatively, if attention is allocated to the three distinguishing features serially but in no particular order (or in parallel), probe detection performance should be similarly high at those three locations compared to the VF location, regardless of SOA. The latter pattern of results would not support the hypothesis of interactive-iterative processing, but would serve as additional support for the hypothesis that attention is directed to distinguishing features during the course of object recognition (Baruch et al., 2014).

5.1. *Method*

5.1.1. *Participants*

Eleven students at the University of Haifa, participated in the experiment. All had normal or corrected to normal vision.

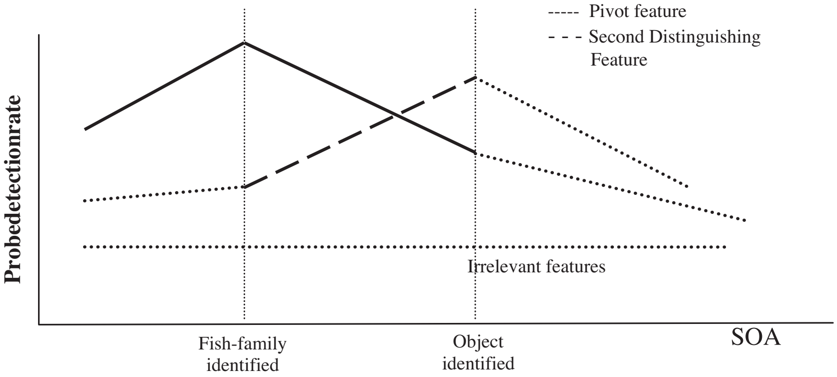

Fig. 5. Predicted probe detection rate in Experiment 1 as a function of probe location and SOA. Probe detection rate is expected to rise initially at the pivot feature until the fish family is identified. It is then expected to decrease at the pivot feature and show a simultaneous increase at the location of the second relevant distinguishing feature (according to fish family) until the specific type of fish is recognized. The detection rate at irrelevant features is expected to remain constant and low. The heavy solid and dashed lines represent the main predictions.

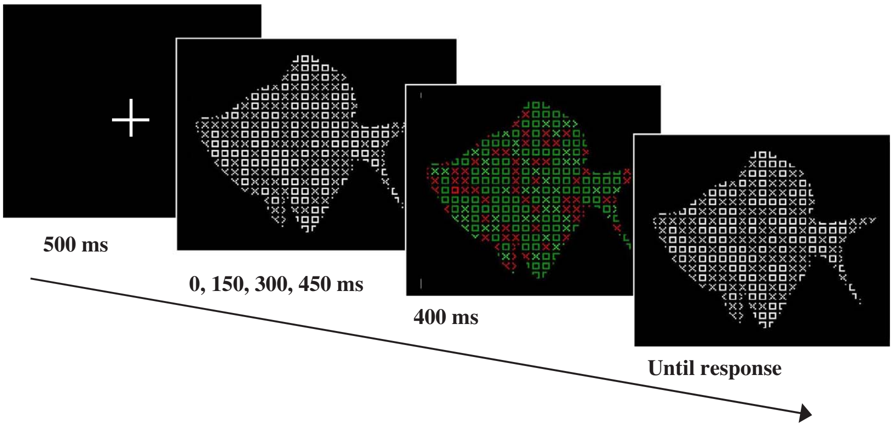

Fig. 6. The sequence of trial events in Experiment 1.

5.1.2. Stimuli

The objects used in this experiment were 2-D textured line drawings of artificial fish, organized into two families, with two types of fish in each family (see Fig. 2). Family membership was distinguished by a single "pivot" feature, the shape of the dorsal fin (DF): *freshwater* fish sharing a *concave* DF and *saltwater* fish sharing a *convex* DF. The two different types of freshwater fish (Grout and Tass) were distinguished further by the shape of their mouth (M), whereas the two different types of saltwater fish (Derch and Fole) were distinguished further by the shape of their tail (T). A fourth feature, the shape of the ventral fin (VF), varied randomly without regard to the type or family of fish. By this scheme, each specific type of fish was defined by the conjunction of values of two different distinguishing features: the pivot feature (DF) and either T or M, depending on fish family (i.e., on the value of DF).

All fish stimuli subtended a visual angle of 8.3° in height and 10° in width. A surface texture was added to the fish by filling their contours with a random array of small 'x' and square elements (each subtending 0.3° height and width, and a 0.1° spacing between the elements). The fish and its texture elements were first presented in gray (on a black background), then, at a selected SOA and for a limited duration (see Design and Procedure), color was added to the texture elements, after which they returned to gray. When colored, each 'x' element was colored either red or green, equally often. On half of the trials (probe-absent trials), all square elements were colored green; on the remaining (probe-present) trials, all square elements were colored green except for one, the *probe*, which was colored red. When present, the probe element appeared equally often at one of the four fish feature locations: M, T, DF or VF. Regardless of the specific shape (value) of the relevant feature, the probe element was presented on the central axis (aligned with the center of the screen/fish), no more than four elements away from the fish shape contour on the x-axis (in the case of M and T) and on the y-axis (in the case of DF and VF).

5.1.3. Design and procedure

The primary task in this experiment was recognition of the presented fish, with each type of fish presented equally often. Each trial began with the display of a small fixation cross at the center of the screen for 500 ms, followed by the display of a gray textured fish on a black background at the center of the screen in either a left or right orientation, counterbalanced across participants. At one of four possible SOAs (0, 150, 300 and 450 ms) the display of the gray fish was replaced by the display of an identical textured colored fish for 400 ms, after which the fish returned to gray and remained until the participant responded (Fig. 6). On 50% of the trials at each SOA, a red-square conjunction probe was present during the colored period, appearing equally often at one of the four fish feature locations. Participants made their recognition response by pressing one of two keys with the right hand for "Grout" or "Tass" and one of two keys with the left hand for "Derch" or "Fole,"[4] double-pressing the key to additionally indicate the detection of a probe. Audio feedback was provided after a recognition error; no feedback was provided for missed or falsely detected probes.

The experiment employed a three-way factorial within-subjects design: 2 (family relationship: saltwater fish/freshwater fish) × 4 (probe location) × 4 (probe SOA). A training phase (see General Method) was administered at the beginning of the experiment to familiarize the participants with the different types of fish, the secondary probe-detection task, the response keys, and the single-press (recognition only) and double-press (recognition + probe detected) response types. Following that, 30 additional practice trials preceded the first block of experimental trials. There were a total of 2048 experimental

[4] As pointed out by a reviewer, the mapping between fish categories and response keys caused the value of the dorsal fin to be informative regarding which hand would be making the response, as well as whether the mouth or tail would be the other relevant distinguishing feature. This may perhaps have added to the optimality of allocating attention first to the dorsal fin.

Table 1
Experiment 1: Mean probe detection rates for the two fish families at the four probe locations as a function of SOA.

	Saltwater fish								Freshwater fish							
	Dorsal Fin		Mouth		Ventral Fin		Tail		Dorsal Fin		Mouth		Ventral Fin		Tail	
SOA (ms)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	74.0	26.6	45.8	28.9	31.4	24.27	58.4	35.1	71.5	22.4	35.5	33.9	27.3	23.4	51.8	33.7
150	77.8	21.9	55.7	29.8	36.6	27.4	52.6	36.8	77.6	20.9	40.4	32.1	36.1	30.1	55.6	34.7
300	69.1	24.4	60.9	23.2	37.7	28.0	52.6	37.4	66.6	21.3	47.2	33.1	35.5	30.1	64.1	33.3
450	64.0	25.9	61.3	27.2	37.2	27.1	53.1	37.6	61.5	24.1	48.8	28.7	44.6	28.7	68.4	24.4

trials, organized into 16 blocks of 128 trials each, with short breaks in between.

Finally, note that the probe duration used in this experiment (400 ms) is rather long relative to the manipulated differences in probe SOA (steps of 150 ms), causing some degree of overlap, particularly between probes presented at “adjacent” SOAs. This large overlap was designed to compensate for individual differences in the time course of attentional shifts (and in the time needed to detect the probe). However, it also means that differences in probe detection rates at the various probe locations as a function of probe SOA can only be used to gauge the *order* of attentional allocations at different stimulus locations, and not the exact timing of these allocations.

5.2. Results and Discussion

5.2.1. Recognition task

Performance in the recognition task was highly accurate, with 88.5% accuracy overall. Recognition accuracy was slightly, though not significantly, higher for saltwater fish (91.5%) than for freshwater fish (85.6%), $F(1, 10) = 2.3$, $p = .16$, $\eta_p^2 = 0.19$, and was completely unaffected by the presence (88.6%) or absence (88.5%) of the probe. Mean RT for correct recognition responses (discarding responses below 200 ms or above 3400 ms; 5% of all trials) tended to be slower for saltwater fish (1503 ms) than for freshwater fish (1418 ms), but this trend was also nonsignificant, $F(1, 10) = 2.4$, $p = .15$, $\eta_p^2 = 0.20$. In view of the opposing nonsignificant trends for recognition latency and accuracy, the slight performance differences between the two fish families suggest a speed-accuracy tradeoff rather than a true difference in recognition difficulty. Importantly, correct recognition responses were faster on probe-absent (1312 ms) than on probe-present (1369 ms) trials, $F(1, 10) = 7.9, p < .05, \eta_p^2 = 0.44$). This result indicates that, in line with the instructions to treat probe detection as a secondary task, the participants were not actively searching for the probes, in which case probe-absent responses should have been slower than probe-present responses (cf. Treisman & Gelade, 1980). Thus, we assume that any disruption of the natural process of fish recognition by the secondary probe-detection task was negligible.

5.2.2. Probe detection

Probe detection rates for each participant were calculated for probe-present trials on which the recognition response was correct.[5] Mean probe detection rates for the two fish families at the four probe locations as a function of SOA are presented in Table 1 and plotted in Fig. 7. The probe detection rates were submitted to a 2 (fish family) × 4 (probe location) × 4 (SOA) repeated measures ANOVA. The analysis revealed significant main effects of probe location, $F(3, 30) = 18.03$, $p < .001$, $\eta_p^2 = 0.64$, and of SOA, $F(3, 30) = 5.04$, $p < .01$, $\eta_p^2 = 0.34$, and significant interactions between probe location and fish family, $F(3, 30) = 8.79$, $p < .001$, $\eta_p^2 = 0.47$), SOA and fish family, $F(3, 30) = 4.16$, $p < .05$, $\eta_p^2 = 0.29$, and probe location and SOA, $F(9, 90) = 6.49$, $p < .001$, $\eta_p^2 = 0.39$. Most importantly, the three-way interaction between fish family, probe location and SOA was significant, $F(9, 90) = 2.41$, $p < .05$, $\eta_p^2 = 0.19$, indicating that the probe detection rate at one or more probe locations differed between the two fish families as a function of SOA. As can be seen in Fig. 7, the overall pattern is in line with the predictions of the optimal, interactive-iterative algorithm. For probes appearing at the DF, the pattern of

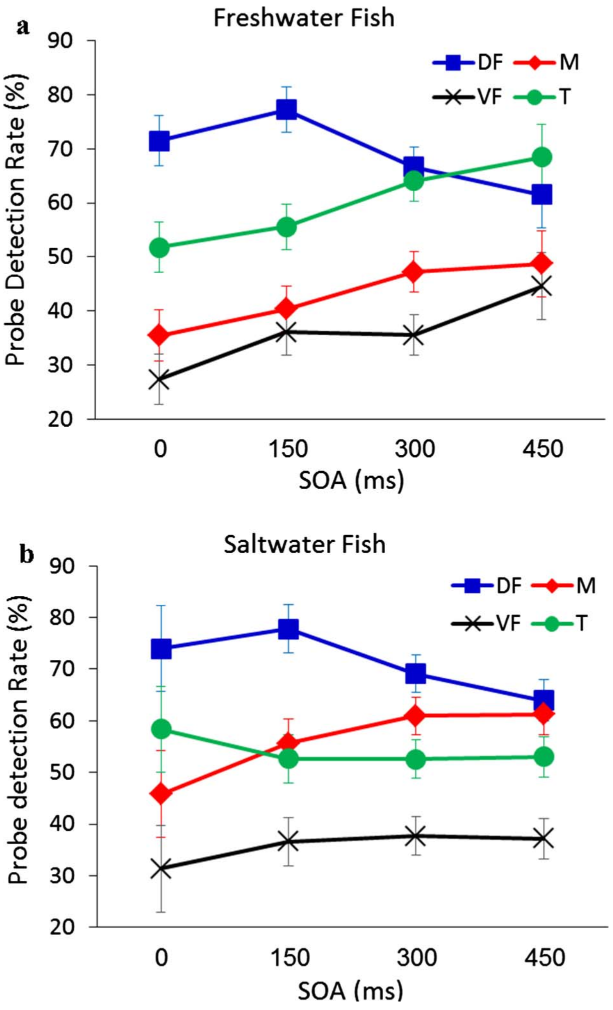

Fig. 7. Results of Experiment 1. Probe detection rates as a function of fish family (panel a – freshwater fish, panel b – saltwater fish), SOA and probe location (DF- dorsal fin, M – mouth, VF – ventral fin, T – tail). For both fish families, the pivot feature was the DF, leading to the initial advantage in probe detection rate at that location. The second relevant distinguishing feature depended on the fish family: tail in the freshwater condition and mouth in the saltwater condition. The VF was irrelevant in both conditions. Error bars represent ± 1 standard error of the within-subject probe-location effect, calculated separately for each SOA and fish family.

[5] Probe false-alarm rate (the percentage of falsely detected probes on probe-absent trials) was very low overall (less than 10% in all experimental conditions; M = 6.5% across all conditions) and there were no significant differences or interactions between conditions on this measure (all $Fs < 1.3$).

detection rate as a function of SOA was similar in both fish families, as expected by its role as the pivot feature, but the pattern of detection rate for probes at the mouth and tail as a function of time differed between the two fish families, as expected by the different role of these two features (diagnostic or undiagnostic) in the different fish families.

Examining the three-way interaction in more detail confirmed that for both fish families the detection rate for probes at the DF displayed a similar pattern of change in detection rates over time: Across the two fish families, detection rate was at an intermediate level (72.8%) at SOA = 0, increasing to its maximum level (77.7%) at SOA = 150 ms, $F(1, 10) = 4.9$, $p < .05$, $\eta_p^2 = 0.33$, followed by a decrease (to 68%) at SOA = 300 ms, $F(1, 10) = 15.2$, $p < .01$, $\eta_p^2 = 0.60$, with a further decrease (to 62.9%) at SOA = 450 ms, $F(1, 10) = 5.8$, $p < .05$, $\eta_p^2 = 0.37$. Moreover, for both fish families, the decrease in detection rate at the later SOAs for probes appearing at the DF was accompanied by a corresponding increase in detection rate for probes appearing at the location of the other relevant distinguishing feature (family dependent: M for saltwater fish, T for freshwater fish). Across the two fish families, detection rate at the family-dependent distinguishing feature was higher at SOA = 300 ms (63.3%) than at SOA = 150 ms (56.1%), $F(1, 10) = 11.47$, $p < .01$, $\eta_p^2 = 0.53$, and there was a further increase between SOA = 300 ms (63.3%) and SOA = 450 ms (64.7%), $F(1, 10) = 19.4$, $p = .001$, $\eta_p^2 = 0.66$. This overall shift of attention from the pivot feature to the family-dependent distinguishing feature over time was also observed in specific contrasts conducted separately for each fish family, as indicated by a significant interaction between probe location (DF vs. M) and SOA for saltwater fish, $F(2, 20) = 7.67$, $p < .01$, $\eta_p^2 = 0.48$, and a corresponding interaction between probe location (DF vs. T) and SOA for freshwater fish, $F(2, 20) = 12.19$, $p < .001$, $\eta_p^2 = 0.55$.

Finally, as expected, for both fish families the detection rate of probes located at the VF was lower than the rate observed at the other (distinguishing) feature locations, at all SOAs. Also as expected, for the saltwater fish there was no significant change in VF probe detection rate as a function of SOA, $F(3, 30) = 1.4$, $p = .25$, $\eta_p^2 = 0.13$. The unexpected increase with SOA observed for the freshwater fish, $F(3, 30) = 5.7$, $p < .05$, $\eta_p^2 = 0.36$, may be a consequence of the relative proximity of the VF and tail shape contours (see Fig. 2), and hence, some small benefit for the detection of VF probes ensuing from the spillover of attention to the tail feature (at the later SOAs) on the freshwater-fish trials.

Another small anomaly in the results is worth noting. On freshwater-fish trials, in addition to the expected increase in probe-detection rate at the later SOAs for probes appearing at the tail (the distinguishing feature that should be examined after discerning the shape of the DF pivot feature), a similar increase in probe-detection rate was also observed for probes appearing at the location of the irrelevant mouth feature (see Fig. 7), though the absolute levels of probe detection were much lower for the mouth (43.0%) than for the tail (59.9%), $F(1, 10) = 13.1$, $p < .01$, $\eta_p^2 = 0.57$. The unexpected increase in probe detection rate at the irrelevant mouth feature may reflect a “noisy” implementation of the optimal algorithm, with the initial allocation of attention to the DF being followed, on a relatively small proportion of trials, by an erroneous shift of attention to the mouth rather than to the tail. If so, this apparently occurred less often when the DF pivot feature indicated that the stimulus was a salt-water fish, as there was no corresponding increase in the detection of probes appearing at the location of the irrelevant tail feature on salt-water fish trials.

In sum, the results of Experiment 1 provide evidence indicating that object recognition may involve, when necessary or expedient, an iterative interaction between bottom-up and top-down processing. The overall pattern of attentional allocations observed in this experiment is consistent with the (somewhat noisy) implementation of an interactive-iterative recognition algorithm, in which the initial allocation of attention to one highly diagnostic “pivot” feature is used to narrow down the hypotheses regarding object identity, and on that basis, shift attention to a further diagnostic feature that can conclusively identify the object. Similar results using eye movements to track overt attention were reported by Blair et al. (2009) in the context of category learning, and the findings of the present experiment can be seen to augment those results: Beyond the difference in the method used to track attention, the stimuli used by Blair et al. (2009) were “amoeba-like” objects spanning a visual angle of 24° (compared to 10° here), and the average time required to categorize each object was on the order of several seconds. It is reasonable to assume that the stimulus sizes and recognition latencies in the present study are more representative of everyday object recognition.

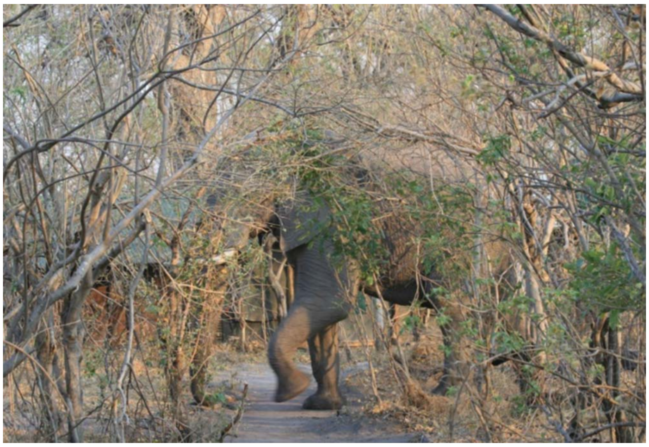

Fig. 8. Example of partial occlusion in a natural scene.

6. Experiment 2

The results of Experiment 1 indicate that object recognition may involve the interactive-iterative allocation of attention to distinguishing features when these features are spatially distributed, and in particular, when the value of one of these features constrains the object hypotheses to a particular object category, and by doing so, determines which of the remaining object features should be examined next. The aim of Experiment 2 is to examine a different general type of situation that similarly calls for an interactive-iterative recognition process.

As discussed earlier, in natural scenes there are many situations in which the initial information that can be extracted quickly and easily from the visual input may be insufficient to conclusively identify a particular object. Such a situation may occur, for example, when the object is partly occluded. If the critical information needed to recognize the object is completely hidden from view, one may need to move one’s head, reposition one’s body, or remove the obstruction in order to improve the amount and quality of relevant visual information. In other cases, sufficiently diagnostic information may remain in the occluded image, but this may not be the information that is typically used as the basis for recognition, and its extraction may require additional scrutiny. To illustrate, if we assume (for the purpose of this example) that an elephant is typically recognized by its overall size and shape, color, and the presence of a trunk, when encountered in its natural habitat some or all of these features may be fully or partly occluded (see Fig. 8). In such a case, although a tentative hypothesis of “elephant” may direct one’s attention, say, to the expected location of the trunk, when this fails to yield the expected diagnostic information, recognition may then need to be based on (and attention redirected to) other, generally less salient features, such as the size and shape of the legs and the texture of the skin.

In the present experiment we examined the idea that different object features are given different priorities in the recognition process, depending on their diagnostic value and perceptual discriminability, and that adjustments to these priorities are made “online” in an interactive-

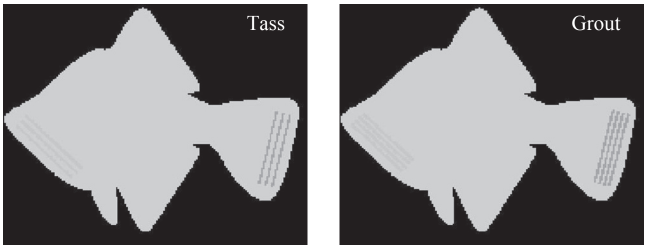

Fig. 9. The fish stimuli used in Experiment 2. The type of fish could be identified based on the number of stripes (3 or 4) comprising either of the two feature patterns—one located at the mouth (low-contrast pattern) and the other located at the tail (medium-contrast pattern).

iterative manner, according to the visual data limitations encountered in specific situations. To that end, the dynamics of the recognition process—and the corresponding dynamics of attentional allocation in that process—were examined in a situation in which one of two equally diagnostic but differentially discriminable distinguishing features was sometimes severely degraded due to partial occlusion.

Two types of fish were used in this experiment (see Fig. 9), which were distinguishable on the basis of either of two local features. One of these features was relatively easy to discriminate (three vs. four medium-contrast stripes at the location of the tail), whereas the other was relatively difficult to discriminate (three vs. four low-contrast stripes at the location of the mouth).

Since the diagnosticity of the two features was identical, and one feature (the medium-contrast stripe pattern at the tail) was more perceptually discriminable than the other (the low-contrast stripe pattern at the mouth), we expected that for the purpose of recognizing each presented fish, attention would be directed primarily to the stripe pattern at the tail. On some of the trials, however, this feature was almost entirely occluded, such that its value (3 stripes or 4 stripes) became very difficult to discriminate—much more difficult than discriminating the number of stripes in the pattern at the mouth (see Fig. 8, panel c). On such trials, we envisioned two alternative attentional strategies that might be used to recognize the fish. If the occlusion of the more discriminable stripe pattern at the tail could be detected immediately, without focused attention (e.g., using low spatial-frequency information; Bar, 2003), then on that basis, participants might simply direct attention from the start to the (unoccluded) low-contrast stripe pattern and use that information to recognize the fish. However, we deliberately designed the occlusion pattern so that it would be difficult to distinguish from the unoccluded pattern on the basis of low spatial frequency information alone. This, together with the fact that the tail pattern was unoccluded and relatively easy to discriminate on two-thirds of the trials (on one third of the trials the mouth pattern was occluded, and on one third neither pattern was occluded), led us to expect that attention would initially be directed to the location of the tail pattern on all trials, but that when this feature was occluded (i.e., when focused attention at the occluded location failed to yield diagnostic information), attention would then be redirected to the generally less discriminable but now more discriminable feature pattern at the mouth. Fig. 10 presents this expected algorithm for the interactive-iterative allocation of attention during the fish recognition process.

As a window to the actual dynamics of attention during the course of recognizing the unoccluded and partly occluded fish, we again used the variable-SOA visual-probe method. These dynamics were expected to be reflected in the probe detection rate as follows: On *no-occlusion* and *mouth-occluded* trials, in which the relatively discriminable feature information at the tail is fully available, attention should be directed initially to the tail and remain there until the fish is recognized. Thus, there should be an early advantage in the detection of probes located at the tail compared to probes located at the mouth, and this difference should increase at longer SOAs, with probe detection at the tail increasing and probe detection at the mouth decreasing until recognition is achieved. However, on *tail-occluded* trials, in which the diagnostic feature information at the tail is severely degraded, assuming that the occlusion can only be detected with focused attention to the tail, attention is again expected to be directed initially to the tail (again yielding initially better probe detection at the tail than at the mouth), but once the occlusion is detected and attention is consequently shifted to the mouth, probe detection at the tail should decrease and probe detection at the mouth should increase until recognition is achieved. Note that if we are wrong in our assumption that focused attention will generally be needed to detect the occlusion of the striped feature pattern at the tail, and on such trials attention is in fact allocated directly to the pattern at the mouth, we would then expect the pattern of probe detection rates at mouth and tail as a function of SOA to be a mirror image of the predicted pattern for the no-occlusion and mouth-occluded trials.

With regard to the actual fish recognition performance, we also expected slower (and perhaps less accurate) recognition responses on

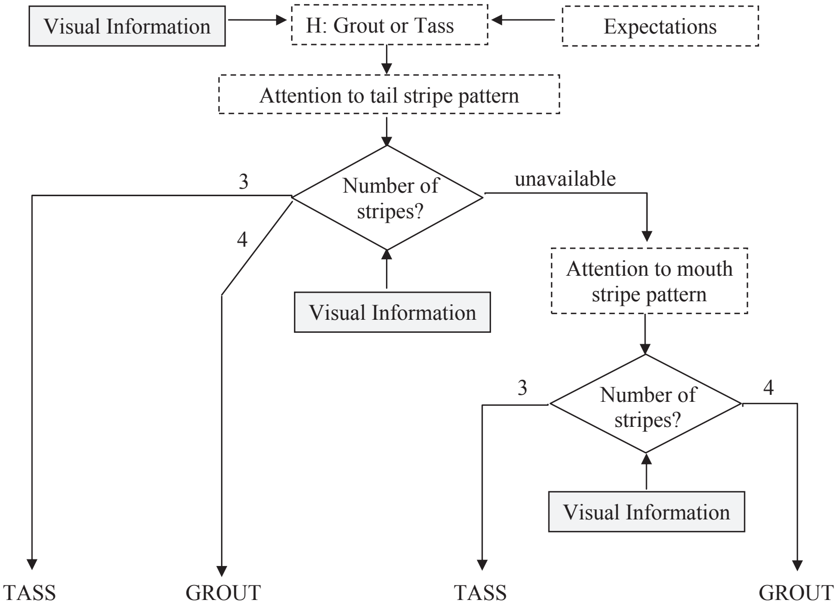

Fig. 10. Schematic diagram of the predicted interactive-iterative fish recognition process in Experiment 2. The initial hypothesis (H: fish is either Grout or Tass) initiates top-down direction of attention to the tail—the location of the generally more discriminable, medium-contrast stripe feature. Visual information extracted at the attended location provides either diagnostic information (i.e., the number of stripes) that enables recognition of the fish, or alternatively, indicates that the stripes are occluded (i.e., diagnostic information is unavailable or highly degraded). In the latter case, attention will be shifted in a top-down manner to the generally less discriminable (but equally diagnostic) low-contrast stripe feature at the mouth of the fish. Note that this algorithm assumes that (a) extracting the number of stripes in either feature pattern requires focused attention, and (b) the partial occlusion of the stripe pattern at the tail (or at the mouth) cannot be detected without focused attention (see underwater scene examples in Fig. 11).

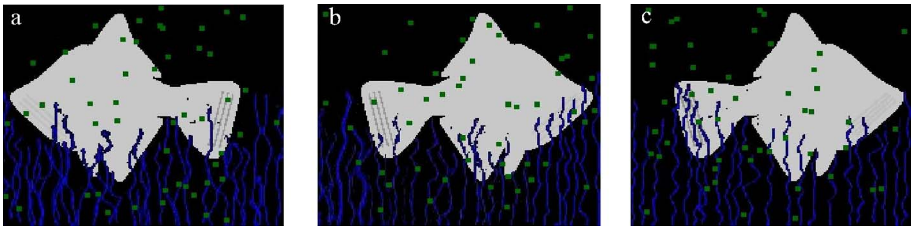

Fig. 11. Examples of the underwater scenes used in Experiment 2: (a) no occlusion of distinguishing features (b) heavy occlusion of the striped feature pattern at the mouth; (c) heavy occlusion of the striped feature pattern at the tail.

tail-occluded trials than on no-occlusion and mouth-occluded trials. Slower responding on tail-occluded trials might simply stem from the lower discriminability of the stripe feature pattern at the mouth, and thus be observed even if attention is initially directed to the mouth on such trials. However, it might also stem from the predicted additional attentional iteration, as attention is shifted first to the tail and then to the mouth once the occlusion of the tail is detected. The probe-detection results will allow us to decide between these two possibilities.

6.1. Method

6.1.1. Participants

Twelve students at the University of Haifa, participated in the experiment. All had normal or corrected to normal vision.

6.1.2. Stimuli

The objects for the recognition task in this experiment were 2-D gray (RGB [200,200,200]) line drawings of artificial fish. There were two types of fish, “Tass” and “Grout,” distinguishable both by the number of stripes (three vs. four, respectively) in a feature pattern located at the tail, and by the (same) number of stripes (three vs. four, respectively) in a feature pattern located at the mouth. The stripes comprising the tail feature appeared in an intermediate-contrast gray (RGB [150,150,150]) whereas the stripes comprising the mouth feature appeared in a low-contrast gray (RGB [190,190,190]), and therefore were more difficult to perceive (see Fig. 9).[6] The two feature patterns were equidistant from the center of the fish.

The fish, subtending 11° × 9°, were embedded in an underwater scene (black background, subtending 13° × 11.3°) that included underwater blue[7] plants (see Fig. 11). The hue and brightness of the plants was varied to create a naturalistic appearance. Three underwater scenes were used. In one scene, the plants partially occluded the stripes located at the mouth of the fish (Fig. 11, panel b), in one scene the stripes at the tail were partially occluded (Fig. 11, panel c), and in one scene there was no occlusion of distinguishing features (Fig. 11a). Note that although the occlusion was not complete, the relevant diagnostic information (number of stripes) in the occluded feature pattern became very difficult, if not impossible, to extract, even under focused attention. In addition to the plants, the underwater scene included 50 “plankton”—small green square elements subtending 0.3° that onset and offset asynchronously for random durations (limited by the refresh rate of the CRT monitor; yielding a “shimmer” effect) at randomly chosen locations across the scene. The brief onsets and offsets were designed to mask the onset and offset of the probe.

The probe in this experiment was also a small green square, the same size as the plankton elements, but in a brighter shade of green (RGB [0,255,0]). The color of the plankton elements was set initially to (RGB [0,100,0]) for all participants, but this value was adjusted dynamically during the training and practice sessions using a titration procedure based on each participant’s probe detection rate, such that the contrast between the plankton elements and the probe was increased when the probe-detection rate in the preceding 10 probe trials was below 40% and was decreased when the probe-detection rate was above 60%. The RGB value at the end of the practice trials was used as the fixed value for all of the subsequent experimental trials. Across the 12 participants, the RGB value of the plankton elements used on the experimental trials varied between RGB (0,70,0) and RGB (0, 120,0), with an average of RGB (0,93,0).

6.1.3. Design and procedure

The primary task in this experiment was recognition of the presented fish, with each type of fish (Grout or Tass) presented equally often. Each trial began with the display of a central fixation cross for 500 ms, which was replaced by the display of a fish embedded in an underwater scene (including the vegetation and small green squares of plankton) at the center of the screen. The fish (and corresponding scene) were presented facing left or right equally often. Scenes representing the three occlusion conditions (tail occluded, mouth occluded, or no occlusion) were presented equally often, randomly intermixed within blocks. Immediately following scene onset, the green plankton squares began to “shimmer” (as described earlier) until the end of the trial. On 50% of the trials, at one of five SOAs (0, 80, 160, 240 or 320 ms), the bright green probe square was displayed for 100 ms, equally often at one of the two probe locations (tail or mouth; the exact probe location was randomly jittered at each location in a small rectangular area subtending 0.2° × 0.4°). Participants made their recognition response by pressing one of two keys, double-pressing the key to additionally indicate the detection of a probe. Audio feedback was provided after a recognition error; no feedback was provided for missed or falsely detected probes.

The experiment employed a three-way factorial within-subjects design: 3 (occlusion condition: no occlusion, tail occluded, mouth occluded) × 2 (probe location: mouth, tail) × 5 (probe SOA). As in the previous experiment, both a training phase (80% probe-present trials) and an additional block of practice trials (50% probe-present trials) were administered at the beginning of the experiment. In this experiment, however, the practice block included 60 trials instead of 30, and both the training and practice trials were used to adjust the contrast between the probe and distractor plankton squares by the titration method, described earlier. There were a total of 1500 experimental trials, organized into 7 blocks of 200 trials each and a last block of 100 trials, with short breaks in between.

6.2. Results and discussion

6.2.1. Recognition task

Performance in the recognition task was highly accurate with 94.88% accuracy overall. A one-way (occlusion condition: none,

[6] One might be concerned that the two feature patterns differed not only in perceptual discriminability but also in visual saliency, in which case attention might be drawn to the tail feature pattern automatically, in a bottom-up manner. To check this possibility, visual saliency maps were produced (using the “Saliency Toolbox”; Walther & Koch, 2006) for 960 randomly selected displays from each occlusion condition (including random vegetation-plankton configurations). The location of the tail feature pattern was identified as the most salient visual area on only 19.4%, 22%, and 17.7% of the maps in the no-occlusion, mouth-occluded, and tail-occluded conditions, respectively. The corresponding percentages for the mouth feature pattern were 21%, 17.1% and 27.9% respectively. Particularly noteworthy is the finding that the location of the tail feature pattern was actually identified as the most salient visual area somewhat less often than was the location of the mouth feature pattern, both in the no-occlusion and in the tail-occluded conditions.

[7] For interpretation of color in Figs. 11 and 13, the reader is referred to the web version of this article.

Table 2
Experiment 2: Mean probe detection rates at mouth and tail locations as a function of SOA for the three occlusion conditions.

	No occlusion				Mouth occluded				Tail occluded			
	Mouth		Tail		Mouth		Tail		Mouth		Tail	
SOA (ms)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0	30.8	15.9	51.9	24.4	24.0	16.6	54.67	20.1	29.3	18.7	39.0	18.6
80	55.5	22.8	67.9	25.5	58.1	28.8	76.85	19.2	65.6	18. 9	62.3	23.4
160	42.1	24.7	70.6	18.2	43.2	24.4	80.6	17.0	71.4	23.0	53.5	25.4
240	28.9	20.00	80.3	13.6	41.7	18.6	76.86	19.4	61.0	24.1	55.5	24.1
320	28.0	17.5	76.2	17.3	17.1	14.1	72.5	23.7	60.5	20.5	48.0	17.5

mouth, tail) repeated measures ANOVA conducted on recognition accuracy data in the no-probe trials, revealed a significant effect of occlusion condition, $F(2, 22) = 8.35$, $p < .01$, $\eta_p^2 = 0.43$. Participants correctly recognized the fish on 95.8% of the trials in the no-occlusion condition, on 95.7% in the mouth-occluded condition and on 93.1% of the trials in the tail-occluded condition. Further analysis showed a significant difference between the tail-occluded condition and both the no-occlusion and mouth-occluded conditions, $F(1,11) = 11.66$, $p < .01$, $\eta_p^2 = 0.52$, and $F(1, 11) = 10.29$, $p < .01$, $\eta_p^2 = 0.48$, respectively. There was no significant difference between the no-occlusion and the mouth-occluded conditions, $F < 1$. As in Experiment 1, recognition accuracy was unaffected by the presence (94.5%) or absence (95.1%) of the probe.[8]

Similarly, analysis of the RT data (discarding responses below 200 ms or above 2000 ms; 4% of all trials) revealed a significant effect of occlusion condition, $F(2, 22) = 138.59$, $p < .001$, $\eta_p^2 = 0.93$). The mean RT was 911 ms in the no-occlusion condition, 904 ms in the mouth-occluded condition and 1139 ms in the tail-occluded condition. Further analysis showed no difference between the no-occlusion condition and the mouth-occluded conditions, $F < 1$. However a highly significant difference was found between the tail-occluded condition and both the no-occlusion and mouth-occluded conditions, $F(1, 11) = 159.89$, $p < .001$, $\eta_p^2 = 0.94$ and $F(1, 11) = 174.16$, $p < .001$, $\eta_p^2 = 0.94$, respectively).

Thus, as expected, the recognition process in the tail-occluded condition was slower and less accurate than in the no-occlusion and mouth-occluded conditions, which did not differ from each other. The finding of equivalent recognition performance in the no-occlusion and mouth-occluded conditions (i.e., regardless of whether the distinguishing feature information was available at the mouth) is consistent with the expectation that the recognition process would be based primarily on the more perceptually discriminable feature information at the tail, whenever that information is available (not occluded). As discussed earlier, slower (and less accurate) recognition in the tail-occluded condition than in the other two conditions could simply reflect the extra time needed to extract the diagnostic information from the low-contrast stripe pattern at the mouth. However, it might also reflect the time needed for an additional attentional iteration, as attention is shifted first to the tail and then to the mouth once the occlusion of the tail is detected. Examination of the probe-detection results will yield further light on this possibility.

6.2.2. Probe detection

Probe detection rate was calculated for all probe-present trials on which the recognition response was correct.[9] Probe detection rates at mouth and tail locations as a function of SOA for the three occlusion conditions are presented in Table 2, and plotted in Fig. 12.

The probe detection rates were submitted first to a 3 (occlusion condition: none, mouth, tail) × 2 (probe location: M, T) × 5 (probe SOA: 0, 80, 160, 240, 320 ms) repeated measures ANOVA. The analysis revealed a significant effect of probe location, $F(1, 11) = 67.1$, $p < .001$, $\eta_p^2 = 0.86$, a significant effect of SOA, $F(4, 44) = 17.56$, $p < .001$, $\eta_p^2=0.62$, and significant interactions between probe location and SOA, $F(4, 44) = 6.05$, $p = .001$, $\eta_p^2 = 0.36$, occlusion condition and SOA, $F(8, 88) = 2.6$, $p < .05$, $\eta_p^2 = 0.19$, and occlusion condition and probe location, $F(2,22) = 51.36$, $p < .001$, $\eta_p^2 = 0.82$. Importantly, a significant three-way interaction was found, $F(8, 88) = 5.43$, $p < .001$, $\eta_p^2 = 0.33$, indicating different patterns of probe detection rates at the different probe locations as a function of SOA in the different occlusion conditions. Inspection of Fig. 12, panels a and b, reveals that, as predicted, the observed pattern in the mouth-occluded and the no-occlusion conditions, in which the relatively discriminable tail feature is not occluded, is very similar, as reflected in the nonsignificant interaction between occlusion condition (no-occlusion vs. mouth-occluded), probe location, and SOA, $F(4, 44) = 1.66$, $p = .11$, $\eta_p^2 = 0.13$. The pattern in these two (tail not occluded) conditions, however, is quite different from the one observed in the tail-occluded condition (Fig. 12, panel c), as reflected in the highly significant interaction between occlusion condition (tail-occluded vs. mean of mouth-occluded and no-occlusion), probe location, and SOA, $F(4, 44) = 9.95$, $p < .001$, $\eta_p^2 = 0.48$.

Given the equivalent patterns observed in the no-occlusion and mouth-occluded conditions, and because our primary interest is in comparing the pattern of attentional allocation in the no-occlusion condition (in which diagnostic object information is available at both feature locations) to the corresponding pattern in the tail-occluded condition (in which the more perceptually discriminable diagnostic information at the tail is temporarily unavailable), we will confine a more detailed analysis of the individual patterns to these latter two conditions.

Beginning with the no-occlusion condition (Fig. 12, panel a), there was a significant main effect of probe location, $F(1, 11) = 139.54$, $p < .001$, $\eta_p^2 = 0.93$, with a higher detection rate at the tail than at the mouth for all SOAs. Interestingly, there was already a probe-detection advantage at the tail at SOA = 0 ms, $F(1, 11) = 9.53$, $p = .01$, $\eta_p^2 = 0.46$. Probe detection rate then increased at SOA = 80 ms, both at the mouth location, $F(1, 11) = 16.83$, $p = .002$, $\eta_p^2 = 0.61$, and at the tail location, $F(1, 11) = 5.35$, $p < .05$, $\eta_p^2 = 0.33$. At this point, probe detection rate at the tail continued to increase with increasing SOA, reaching a maximal value at SOA = 240 ms and then leveling off—an increase that was mirrored by a corresponding decrease in

[8] However, also as in Experiment 1, correct recognition responses were significantly faster (by 51 ms) in the probe-absent trials compared to the probe-present trials, $F(1, 11) = 42.91$, $p < .001$, $\eta_p^2 = 0.80$. Once again, this difference is the opposite of what would be expected if the participants had been actively searching for the probes (cf. Treisman & Gelade, 1980), indicating that the participants were in fact following the instructions to treat fish recognition as their primary task.

[9] As in Experiment 1, probe false-alarm rate (FAR) was very low overall (less than 11% in all experimental conditions; M = 8.2% across all conditions) and there were no significant differences or interactions between conditions on this measure. There was a nonsignificant trend for higher FAR at the tail (8.6%) than at the mouth (7.8%), $F(1, 11) = 3.7$, $p = .08$, $\eta_p^2 = 0.25$; all other Fs < 1.7).

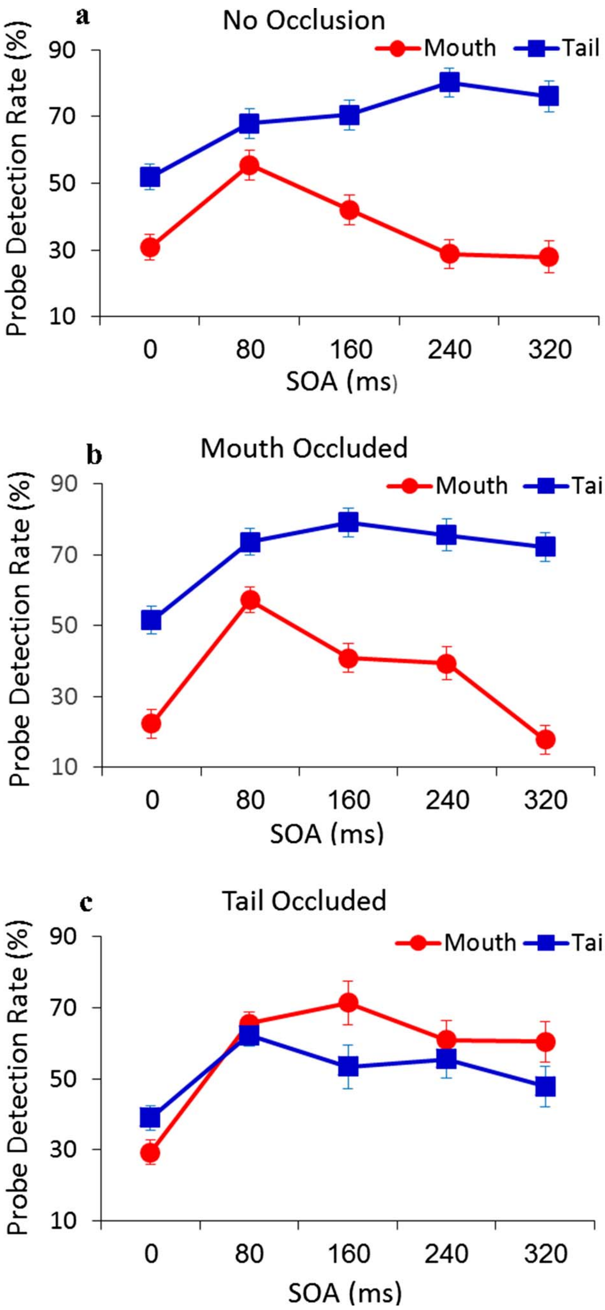

Fig. 12. Results of Experiment 2. Probe detection rates as a function of occlusion condition (panel a – no occlusion, panel b – mouth occluded, panel c – tail occluded), SOA and probe location (mouth or tail). Error bars represent ± 1 standard error of the within-subject probe-location effect, calculated separately for each SOA and occlusion condition.

detection rate at the mouth, reaching its minimum value at SOA = 240 ms, before leveling off. These opposing changes in probe detection rate at the tail and mouth after 80 ms are reflected in a significant 2-way interaction between probe location and SOA (80, 160, 240, 320 ms), $F(1, 11) = 8.40$, $p < .001$, $\eta_p^2 = 0.43$.

The finding of a probe detection advantage at the tail at SOA = 0 ms suggests that on some of the trials, by 100 ms (the probe duration) the location (left or right) of the higher-contrast feature pattern at the tail had been determined and focused spatial attention had reached that location. Yet, the parallel increase in detection rate at both the mouth and tail locations between SOA = 0 ms and SOA = 80 ms requires explanation. One possibility is that attention was initially divided between both ends of the fish stimulus until the location of the higher contrast tail feature could be determined, and that on the whole, detection of a probe onsetting 80 ms after fish-stimulus onset was easier than detection of a probe appearing together with the fish. By this account, the observed advantage in probe detection at the tail at both of these initial SOAs reflects trials in which the determination of the location of the higher-contrast tail feature, and the focusing of attention on that feature, occurred relatively quickly. An alternative possibility is that on some proportion of trials, attention was initially directed to the mouth rather than to the tail. This might stem from a mistake in determining the location of the higher contrast tail feature on some of the trials, or from some amount of arbitrariness in the choice of which end of the fish to initially attend. By this account, the observed advantage in probe detection at the tail at the two shortest SOAs would reflect a more frequent initial allocation of attention to the tail than to the mouth, with the increase in this difference at longer SOAs indicating that when attention was initially allocated to the tail it was maintained there until recognition, whereas when initially allocated to the mouth, on some proportion of the trials attention was subsequently redirected to the tail. Overall, this second account of the results, which includes an unforeseen attentional iteration on some of the trials (shift from mouth to tail when the more discriminable tail feature is initially mislocated), implies a "noisy" implementation of the expected algorithm. Note that by both accounts, because the advantage in probe detection at the tail was also observed at SOAs that are presumably too short for eye movements, this indicates that covert (as well as overt) attention was involved in the extraction of diagnostic feature information at the tail.

Turning now to the tail-occluded condition (Fig. 12, panel c), it can be seen that even with no diagnostic information available at that location, there is still an initial tendency toward higher probe detection at the tail than at the mouth at SOA = 0 ms, $F(1, 11) = 3.37$, $p = .09$, $\eta_p^2 = 0.23$, though the difference here only approaches significance (by a two-tailed test; it is significant by a one-tailed test). Compared to the corresponding difference at SOA = 0 ms observed in the no-occlusion condition, the difference here is numerically (but not significantly) smaller, $F(1, 11) = 2.48$, $p = .14$, $\eta_p^2 = 0.18$, for the interaction, stemming from a lower detection rate at the tail in the tail-occluded (39.0%) than in the no-occlusion (51.9%) condition, $F(1, 11) = 3.66$, $p = .08$, $\eta_p^2 = 0.25$. Once again, in the tail-occluded condition there was a substantial increase in detection rate between SOA = 0 ms and SOA = 80 ms, both at the tail, $F(1, 11) = 18.21$, $p = .001$, $\eta_p^2 = 0.62$, and at the mouth, $F(1, 11) = 32.06$, $p < .001$, $\eta_p^2 = 0.75$. Unlike in the no-occlusion (and mouth-occluded) conditions, however, here there was no difference in probe detection rates at the tail and mouth locations at SOA = 80 ms, $F < 1$, for the mean difference; $F(1, 11) = 4.89$, $p < .05$, $\eta_p^2 = 0.31$, for the interaction with occlusion condition. Compared to the no-occlusion condition, this change in the tail-occluded condition at SOA = 80 ms stemmed both from a lower detection rate at the tail (by 5.6%), and from a higher detection rate at the mouth, (by 10.2%), though neither of these differences individually reached statistical significance.

The lower tail probe detection rates at the two shortest SOAs in the tail-occluded compared to the no-occlusion condition suggests either that contrary to our assumption, the occlusion at the tail could sometimes be detected without focusing attention to the tail (perhaps the primary cause of the reduction at SOA = 0), or that approximately 180 ms (80 ms SOA + 100 ms probe duration) was sometimes sufficient time to allocate attention initially to the tail, detect the occlusion, and redirect attention to the mouth. Nevertheless, it should be noted that at these short SOAs, the probe detection rate at the tail in the tail-occluded condition (51%) was higher than the corresponding detection rate at the mouth in the mouth-occluded condition (41%), $F(1,11) = 10.28$, $p < .01$, $\eta_p^2 = 0.48$. This relatively high rate of probe detection at the tail at short SOAs in the tail-occluded condition cannot be explained simply in terms of initially divided attention, or the arbitrary or mislocated allocation of focused attention—factors that should be equivalent in the tail-occluded and mouth-occluded conditions. Instead, it appears that on a substantial proportion of the trials, the occlusion of the tail pattern was not detectable without focused attention, so that attention was initially directed to the tail on these trials in a top-down manner, before being redirected to the mouth once the occlusion was discovered. Again, because these differences were observed at SOAs that are too short for eye movements, they indicate that covert focused attention could be allocated and reallocated very quickly, in an iterative manner, in response to occlusion of the generally more discriminable tail feature.

The tendency to direct or redirect attention to the low-contrast feature pattern at the mouth when the generally more discriminable information at the tail is temporarily unavailable can also be seen in the divergent patterns of increasing detection of probes at the mouth location and decreasing detection of probes at the tail location between SOA = 80 ms and SOA = 160 ms, $F(1,11) = 5.12, p < .05, \eta_p^2 = 0.32$, for the two-way interaction. This is precisely the opposite of the pattern observed in the no-occlusion (and mouth-occluded) conditions, $F(1,11) = 9.01, p < .05, \eta_p^2 = 0.45$, for the three-way interaction. By SOA = 160 ms, probe detection rate was higher at the mouth than at the tail, $F(1,11) = 4.24, p = .06, \eta_p^2 = 0.28$ ($p < .05$ by a one-tailed test), remaining higher across the two longer SOAs, $F(1,11) = 1.46, p = .25, \eta_p^2 = 0.18$.

Interestingly, however, probe detection rates at the tail also remained relatively high across the interval between SOA = 160 ms and SOA = 320 ms, in comparison to the corresponding detection rates at the mouth in the no-occlusion (and mouth occluded) condition, $F(1,11) = 33.83, p < .001, \eta_p^2 = 0.76$. Conversely, probe detection rates at the mouth were substantially lower at the two longest SOAs (240 ms and 320 ms) compared to the corresponding detection rates at the tail in the no-occlusion (and mouth occluded) condition, $F(1,11) = 7.89, p < .05, \eta_p^2 = 0.42$. These complementary differences suggest a tendency to continue to allocate some attention to the location of the tail feature, presumably after the occlusion of that feature had already been detected. This tendency might reflect some degree of perseverance in attempting to extract the occluded (severely degraded) information from what was generally the preferred feature location.

In sum, the results of this experiment again support the basic proposal that object recognition involves, when needed or expedient, an interactive and iterative allocation of attention to distinguishing features. They provide an important generalization of the results of Experiment 1 (and those of Blair et al., 2009), showing that such a process occurs not only when the objects to be recognized have a nested categorical structure, but also in response to dynamic changes in data quality that affect the availability and perceptual discriminability of the diagnostic information that can be extracted in particular situations. Moreover, the finding that occlusion condition differentially affected the detection of visual probes appearing at the tail and mouth locations at the two shortest SOAs (0 ms and 80 ms), which are too short for eye movements, clearly indicates that these effects were mediated by differences in the spatiotemporal dynamics of covert attention.

On the whole, the participants in this experiment allocated attention first to the tail of the fish, where the diagnostic information was generally easier to extract. When this strategy failed because the tail feature was occluded, attention was redirected to the mouth of the fish where the extraction of diagnostic information was generally more difficult. Thus, information (or lack thereof) extracted during the allocation of attention to one distinguishing feature indicated the need for a subsequent allocation of attention to a different distinguishing feature. It may be that on some of the trials the occlusion of the tail feature was detected initially under divided attention, so that attention could be allocated directly to the less discriminable feature at the mouth (without first focusing on the tail). If so, this would imply that on some of the trials, the allocation of attention during the recognition process was interactive and iterative (visual stimulus information extracted under divided attention was used to direct the subsequent extraction of visual information under focused attention), but the need for an iterative *shift* of focused attention on such trials could be avoided. In any case, as in Experiment 1, the participants appear to have been rather efficient in finding and using an algorithm that minimizes the number of required attentional allocations and the effort needed to extract the diagnostic information under each allocation.

7. General discussion

The present work promotes the view of object recognition as an interactive-iterative process in which attention to distinguishing features plays a crucial role. The experiments were designed at the operational level using a relatively small number of experimentally defined object types, thereby gaining a large degree of control over the object features—both distinguishing and non-distinguishing. Such tasks are typically treated as object categorization tasks, and the lessons learned using such tasks are often directed specifically to the categorization and category learning literatures (e.g., Blair et al., 2009; Rehder & Hoffman, 2005a, 2005b). A basic tenet of the framework advanced here, however (see also Bruner, Goodnow, & Austin, 1962; Grill-Spector & Kanwisher, 2005; Meuwese, Loon, Lamme, & Fahrenfort, 2014; Palmeri & Gauthier, 2004; Schyns & Rodet, 1997; Wagar & Dixon, 2005; Walther & Koch, 2007), is that object recognition is essentially and fundamentally a matter of categorization—a process of discriminating between probable alternatives, with the set of such alternatives for any particular visual stimulus constrained by one's expectations in a specific context. Indeed, both object recognition and object categorization tasks, and all computational models of object recognition, involve identifying a presently viewed visual stimulus as an instance of a particular object type or category. For example, as explained by Walther & Koch, 2007, "by object recognition we mean our visual system's ability to infer the presence of a particular object or member of an object category from the retinal image" (p. 57). This essential equivalence suggests a need for greater cross-talk between the object recognition and category learning literatures (e.g., Palmeri & Gauthier, 2004).

A second basic tenet of the framework, discussed earlier, is that the initial information extracted from the visual scene in a data-driven (bottom-up) manner regarding any particular object is inherently partial. In natural scenes, the far sides of objects are hidden from view and surfaces may undergo occlusion; viewing conditions may be poor, and the information that is diagnostic of object identity may be subtle and spatially distributed. Even under optimal viewing conditions, the initial extracted information may be relatively coarse (carried by low spatial frequencies, e.g., Bar, 2003; Fabre-Thorpe, 2011). Although, depending on context and expectations, the initially extracted information may sometimes suffice for recognition "at a glance" (e.g. Hochstein and Ahissar, 2002), in other cases object recognition will require additional processing, and take longer to complete.

Two experiments were conducted in which, by design, the critical information needed to recognize the object could not be acquired all at once. Such conditions were created using (a) a nested hierarchy of spatially distributed distinguishing features (Experiment 1), and (b) partial occlusion (Experiment 2). In Experiment 1, the general pattern of attentional allocation as a function of time and location was consistent with the idea that attention is first allocated to the object feature with highest (expected) diagnostic value, and then, on the basis of the extracted information, reallocated to the next diagnostic feature in an iterative manner. A similar overall pattern was observed in Experiment 2: Attention was initially directed to the location where the most discriminable diagnostic information was expected. In some cases, however, this information was heavily occluded and consequently attention was redirected to the other diagnostic (but generally less discriminable) feature to accomplish the recognition task.

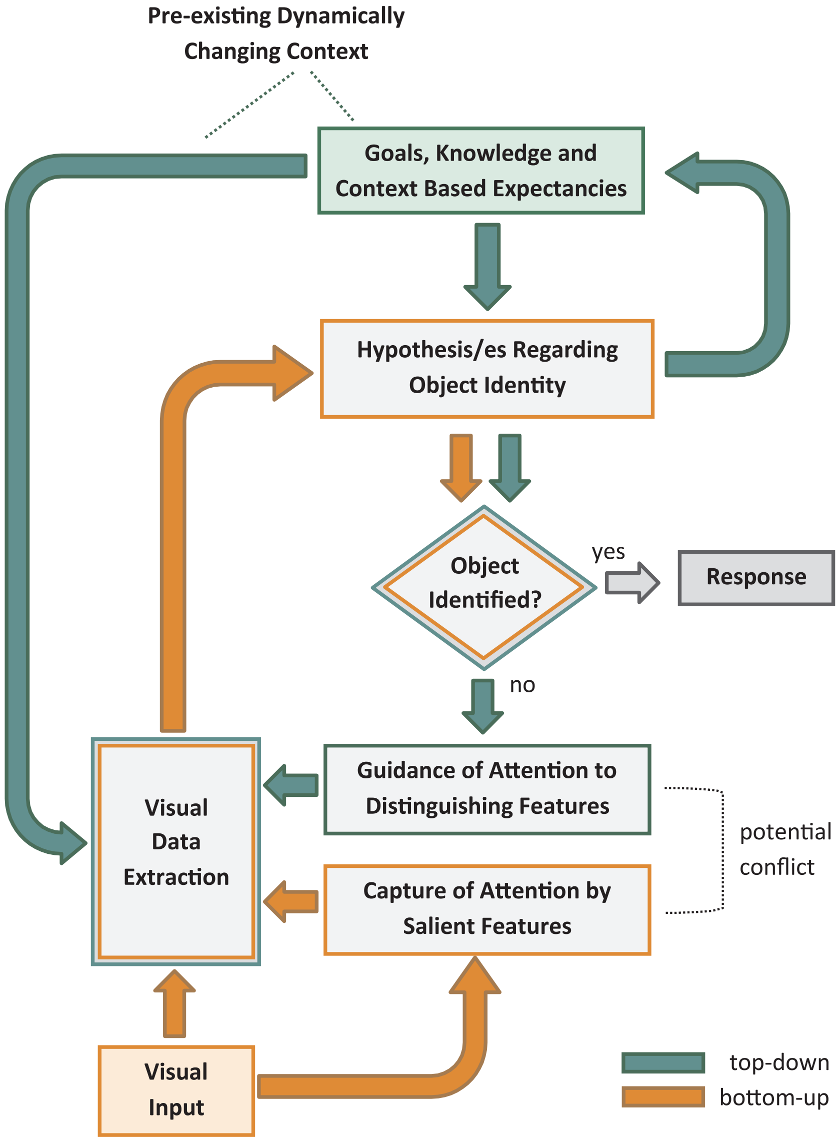
Pre-existing Dynamically Changing Context
Goals, Knowledge and Context Based Expectancies
Hypothesis/es Regarding Object Identity
Object Identified?
yes
Response
no
Visual Data Extraction
Guidance of Attention to Distinguishing Features
Capture of Attention by Salient Features
potential conflict
Visual Input
top-down
bottom-up

Fig. 13. A schematic depiction of the proposed integrative framework for object recognition.

7.1. Interactive-Iterative attentional allocation in object recognition and categorization

Taken together, the results summarized above support the idea that in recognizing objects, attention is directed to distinguishing features[10] in an interactive and—when needed—iterative manner, depending both on top-down expectations and on bottom-up constraints stemming from the availability and quality of stimulus information. That, of course, is not to say that top-down attention is needed for object recognition in all situations, or that all top-down contributions to object recognition are attentional. In this study, we emphasized the role of attention to distinguishing features, and in particular, the potential need, under certain conditions, for an iterative process of attention to distinguishing features —a role that has been relatively neglected. Nonetheless, there are undoubtedly cases in which recognition is achieved very quickly and easily, perhaps too quickly for top-down attention to be deployed (e.g., Li et al., 2002; Thorpe et al., 1996; but see Evans & Treisman, 2005). How should such cases be reconciled? Should we adopt one specific model of the recognition process for some situations, and a different specific model for others? We wish to promote a more parsimonious approach.

Fig. 13 presents a schematic "flow diagram" that instantiates the principles of the proposed integrative framework, described earlier. Within this framework, observers never confront a perceptual event without context-based expectations and world knowledge. These, together with the visual input (whose extraction and analysis are also susceptible to top-down influences[11]), jointly constrain the plausible hypotheses that need to be considered. When these interacting constraints are sufficient to converge on a single hypothesis regarding object identity, the recognition process may terminate quickly and immediately on the first pass (cf. Li et al., 2002; Thorpe et al., 1996). When the initial constraints are not sufficient to converge on a single "winning" hypothesis, the number of remaining hypotheses that satisfy the constraints should nevertheless be quite small. These hypotheses

[10] It should be noted that in both experiments, all distinguishing features were local features (pertaining to only a relatively small part of the fish). However, distinguishing features can also be global or configural features such as texture or shape (see note 1 earlier). Evidence of attention to such features during object recognition cannot be gained using the spatial probe method, but can be obtained using other methods (see Baruch et al., 2014).

[11] Top-down influences on the extraction and analysis of the visual data itself (e.g., Maunsell & Treue, 2006; Summerfield & de Lange, 2014), signified by the left-most top-down arrow in Fig. 13, are discussed further below.

might be expressed as the activation of internal representations of candidate objects, giving special weight to features that discriminate between the competing hypotheses (i.e., distinguishing features). Attention is then directed to these distinguishing features in order to facilitate (further) the extraction of the relevant visual data.[12] Additional iterations will be invoked (with attention guided to the relevant distinguishing features), as needed, until one of these hypotheses can be confirmed.

We believe that this framework is of value in providing a parsimonious integration of the types of processing that underlie the entire gamut of recognition situations: so-called "instantaneous" recognition at one extreme, very effortful time-consuming recognition at the other extreme, and the range of situations that lie between. Rather than a dichotomous conception in which the recognition process is assumed (incorrectly, in our view) to be entirely bottom-up/feed-forward under some conditions but interactive and attentional under others, within this framework object recognition *always* involves an interaction between top-down and bottom-up processing, though the nature and relative contribution of each may vary greatly across different recognition situations.

Some key aspects of this framework—the activation of hypotheses regarding object identity and the testing of these hypotheses via the iterative allocation of visual attention to distinguishing features—are also core components of a theory of object identification put forward by Ganis and Kosslyn (2007). One difference between Ganis and Kosslyn's (2007) proposal and our own, however, concerns the role of context in determining the features that are relied upon in the object recognition process. In their proposal, the features that are attended to (and also primed) in the course of recognizing an object are those that are distinctive of the object representation that best matches the visual input; the set of plausible competing alternatives has no effect. The hypothesis testing process is essentially confirmatory. In contrast, as discussed earlier, in our view object recognition always occurs in context, such that the set of expected plausible alternatives may influence the diagnosticity of different potentially distinguishing features. For example, a feature that is diagnostic for identifying a wolf in a herd of sheep may be of no use at all for identifying a wolf in a herd of wild dogs. Thus, the hypothesis testing process can be characterized as deciding between competing alternatives rather than as confirming or refuting a single most likely hypothesis. Evidence for the context sensitivity of distinguishing features was provided by Baruch et al. (2014): In categorizing different sets of objects, attention was allocated to different features of the same object, depending on the diagnosticity of these features with respect to the overall stimulus set (for further evidence of context-sensitive object representation, see Schyns & Rodet, 1997; Wagar & Dixon, 2005).

Context sensitive allocation of attention to diagnostic features is also important with regard to the potential scalability of such a recognition process. In natural situations in which there is a large number of potential object categories, it is very unlikely that there will be one or two critical features that distinguish a particular object candidate from all other possible object categories, as is assumed by Ganis and Kosslyn (2007). In contrast, as explained earlier, we assume that the initial interaction between top-down expectations and bottom-up visual data substantially constrains the set of relevant object categories, correspondingly reducing the number of relevant distinguishing features that must be examined. If further iterations are needed, the set of relevant categories remaining after each iteration should quickly converge to one (i.e., the object category that is ultimately chosen).

Interestingly, context sensitive allocation of attention to diagnostic features, conceptualized as changes in dimension weights, plays an important role in all major theories of categorization (e.g., Bruner et al., 1962; Medin & Schaffer, 1978; Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961). In many classic category-learning models, there is a single set of attentional weights for a particular task (Kruschke, 1992; Love, Medin, & Gureckis, 2004; Minda & Smith, 2002; Nosofsky, 1986), an assumption which has been referred to as "task-specific" attention (Blair et al., 2009). Attention is deployed to the most relevant features for performing the task in the context of its particular stimulus set. Other theories of categorization, however, allow dimension weights to be tailored not only for the task as a whole, but for particular stimuli as well—an assumption referred to as "stimulus-specific" or "stimulus-responsive" attention (Aha & Goldstone, 1992; Blair et al., 2009; Kruschke, 2001). Whereas task-specific attention implies top-down (task-driven) attention to distinguishing features, stimulus-responsive attention allows both top-down (task-driven) and bottom-up (stimulus-driven) influences on the allocation of attention to occur in an iterative manner. Blair et al. (2009) argued that the primary motivator of stimulus-responsive attention is efficiency. As discussed earlier, using stimuli with a nested hierarchical structure similar to those used in Experiment 1 here, they found systematic temporal patterns in the shifting of eye fixations, such that information gleaned by previous fixations on distinguishing features dictated which subsequent distinguishing feature would be most informative, and therefore fixated next. As illustrated in Experiment 2 here, however, stimulus-responsive attention is not just a way of increasing the efficiency of the process: In many cases one may have no choice but to reallocate attention to a different diagnostic feature when the initial allocation is not successful in yielding sufficiently diagnostic information (e.g., because of occlusion).

7.2. Interactive-Iterative processing in the visual system

We now turn to the broader implications of the present work regarding the nature of visual processing. Debate concerning the role of top-down (knowledge driven) and bottom-up (stimulus driven) processing in visual perception has a long history. Building on von Helmholtz's (1867) notion of "unconscious inference," the constructivist approach to perception (e.g., Epstein, 1973; Gregory, 1966; Hochberg, 1964; Rock, 1983) has emphasized the role of prior knowledge and "intelligent" thought-like processes in allowing the perceiver to resolve the presumed inherent ambiguity of visual stimulation. In strong opposition to this view, Gibson (1966, 1979) put forward the notion of "direct" perception, positing that all of the information needed for unequivocal perception of the environment is available in the visual input (in the form of higher order invariants), so that there is no need for additional perceptual processing beyond the direct "pick up" of this information. Although this debate has never been entirely resolved (see Norman, 2002), one outcome has been greater acknowledgment of the potential richness of the input to the visual system.

In parallel, building on the landmark findings of Hubel and Wiesel (1962, 1968) regarding the hierarchical organization of the visual system, a great amount of work in computational vision and neuroscience has tested the viability of a strictly bottom-up approach in attempting to determine how far perceptual processing can go based on a "feed-forward" analysis of the visual input alone (e.g., Marr, 1982; Reisenhuber & Poggio, 1999). Yet, advances in our understanding of the organization of visual processing in the brain (e.g., Albright & Stoner, 2002; Nakamura, Gattass, Desimone, & Ungerleider, 1993; Tucker & Fitzpatrick, 2003), together with limited success in strictly bottom-up modeling (see Kveraga et al., 2007),[13] has led to increasing acknowledgement of the need for interactive processing in vision

[12] Note that attention can be captured by salient features that are not necessarily the distinguishing features, in which case a conflict may arise between bottom-up and top-down attentional control. In an unpublished study, we found that when the saliency and diagnosticity of object features were placed in competition, recognition latencies increased, but visual probe detection was still higher at the location of the distinguishing feature than at the location of the salient (or any other) nondistinguishing feature. That is, in the competition for the allocation of attention, the distinguishing feature "won".

[13] Interestingly, in the categorization literature it appears that it is the inadequacy of a strictly top-down approach, embodied in the "task-specific" attention models, that has motivated the development of interactive models (Blair et al., 2009).

generally, and in object recognition and categorization in particular (e.g., Bar, 2003; Bullier, 2001; Ganis et al., 2007; Humphreys et al., 1997; Lee, 2002; O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013; Schendan & Maher, 2009; Schendan & Stern, 2008; Ullman, 1995; Wyatte, Jilk, & O'Reilly, 2014).

Most of these proposals include mechanisms by which top-down predictions bias the visual data extraction and analysis (left-most green arrow in Fig. 13) and constrain the probable interpretations of the input (top-center green arrow in Fig. 13). Such interaction between top-down and bottom-up processing may be implemented either in a single feedback pass (e.g., Bar, 2003) or through recurrent processing (e.g., O'Reilly et al., 2013; Wyatte, Jilk, & O'Reilly, 2014). Our work joins these in highlighting the interactive nature of visual processing, but with two special emphases. First, in most other discussions of interactive processing (including Ganis & Kosslyn's, 2007, iterative attentional theory of object identification, discussed earlier), top-down influences are conceived as feedback that is triggered by, and therefore as following, the initial bottom-up analysis of the visual data. By contrast, in our view, visual processing is cyclic (cf. the "perceptual cycle" put forward by Neisser, 1976), and if anything, object recognition actually starts at the top: visual top-down expectations exist, and therefore begin to exert their influence on the recognition of a particular visual stimulus, even before that stimulus is encountered. Such influences are clearly seen, for example, in results showing that priming by category names substantially improves object identification (Reinitz et al., 1989), that objects are recognized better in expected than in unexpected contexts (e.g., Bar & Ullman, 1996; Biederman, 1972, 1981), and that the allocation of attention to distinguishing features is context-dependent (Baruch et al., 2014).

Second, although we agree that top-down biasing of the visual analysis may play a significant role in object recognition, a different, relatively neglected aspect of interactive processing was emphasized in the present work, namely, the iterative allocation of visual attention to distinguishing object features. Apart from this work, and the two exceptions mentioned earlier (Blair et al., 2009; Ganis & Kosslyn, 2007), other proposals that have incorporated interactive-iterative allocation of attention have used it as a means of selecting a specific object in a cluttered scene, and not as part of the recognition process per se (e.g., Deco & Zihl, 2001; Hamker, 2006; Rybak, Gusakova, Golovan, Podladchikova, & Shevtsova, 1998; Schill, Umkehrer, Beinlich, Krieger, & Zetzsche, 2001).

7.3. Concluding remarks

This study revived classic ideas (e.g., Gregory, 1966; von Helmholtz, 1867) regarding the interaction of top-down and bottom-up processes in perception, incorporating them into a framework that explicates the interactive-iterative nature of the process of object recognition, and the role of attention in that process. Although similar ideas have been proposed and discussed by others, a major aim of the present article was to integrate these ideas into a single coherent and parsimonious framework, and to provide empirical evidence supporting a key aspect of this framework (the iterative allocation of attention to distinguishing features) that has been relatively neglected (see also Baruch et al., 2014). More generally, this article joins a growing number of studies emphasizing the inherently interactive nature of the processing that is carried out in the human visual system.

Acknowledgement

This research was supported by grant 1110/10 from the Israel Science Foundation to Ruth Kimchi and Morris Goldsmith, and comprises part of Orit Baruch's Ph.D dissertation, submitted to the University of Haifa. Facilities for conducting the research were provided by the Institute of Information Processing and Decision Making and by the Max Wertheimer Minerva Center for Cognitive Processes and Human Performance.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.10.007.

References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. *Proceedings of the fourteenth annual conference of the cognitive science society* (pp. 534–539). Bloomington, IN: Lawrence Erlbaum.

Albright, T. D., & Stoner, G. R. (2002). Contextual influences on visual processing. *Annual Review of Neuroscience, 25*, 339–379.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience, 15*, 600–609.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*, 617–629.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., ... Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences USA, 103*, 449–454.

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception, 25*, 343–352.

Baruch, O., Kimchi, R., & Goldsmith, M. (2014). Attention to distinguishing features in object recognition. *Visual Cognition. 22*, 1184–1215.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*, 77–80.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy, & J. R. Pomerantz (Eds.). *Perceptual organization* (pp. 213–253). Hillsdale, NJ: Erlbaum.

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*, 1196–1206.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1962). *A study of thinking.* London: Wiley.

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience, 22*. http://dx.doi.org/10.3389/fnhum.2010.00025.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews, 36*, 96–107.

Deco, G., & Zihl, J. (2001). A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system. *Journal of Computational Neuroscience, 10*, 231–253.

Enns, J. T., & Llreas, A. (2008). What's next? New evidence for prediction in human vision. *Trends in Cognitive Sciences, 12*, 327–333.

Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences, 105*, 14298–14303.

Epstein, W. (1973). The process of "taking-into-account" in visual perception. *Perception, 2*, 267–285.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology, Human Perception and Performance, 3*, 1476–1492.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Perception Science, 2* doi:10.3389.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 360*, 815–836.

Friston, K. J., & Kiebel, S. J. (2009). Cortical circuits for perceptual inference. *Neural Networks, 22*, 1093–1104.

Ganis, G., & Kosslyn, S. M. (2007). Multiple mechanisms of top-down processing in vision. In S. Funahashi (Ed.). *Representation and brain* (pp. 21–45). Tokyo: Springer Verlag.

Ganis, G., Schendan, H. E., & Kosslyn, S. M. (2007). Neuroimaging evidence for object model verification theory: Role of prefrontal control in visual object categorization. *Neuroimage, 34*, 384–398.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gillebert, C. R., Op de Beeck, H. P., Panis, S., & Wagemans, J. (2009). Subordinate categorization enhances the neural selectivity in human object-selective cortex for fine shape differences. *Journal of Cognitive Neuroscience, 21*, 1054–1064.

Gregory, R. L. (1966). *Eye and brain.* New York: World University Library.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition as soon as you know it is there, you know what it is. *Psychological Science, 16*, 152–160.

Hamker, F. H. (2006). Modeling feature-based attention as an active top-down inference process. *BioSystems, 86*, 91–99.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wakesleep" algorithm for unsupervised neural networks. *Science, 268*, 1158–1161.

Hochberg, J. (1964). *Perception.* Englewood Cliffs, NJ: Prentice-Hall.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*, 791–804.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interactions and functional architecture of the cat's visual cortex. *Journal of Physiology (London), 160*, 106–154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology, 195*, 215–243.

Hughes, H. C., Nozawa, G., & Kitterle, F. (1996). Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience, 8*, 197–230.

Humphreys, G. W., Riddoch, J., & Price, C. (1997). Top-down processes in object identification: Evidence from experimental psychology, neuropsychology and functional anatomy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 352*, 1275–1282.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 604–610.

Kosslyn, S. M. (1994). *Image and brain.* Cambridge: MIT Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kruschke, J. K. (2001). The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1385–1400.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top–down predictions in the cognitive brain. *Brain and Cognition, 65*, 145–168.

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

Lee, T. S. (2002). Top-down influence in early visual processing: A Bayesian perspective. *Behaviors and Physiology, 77*, 645–650.

Li, F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Science, USA, 99*, 9596–9601.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature, 382*, 616–618.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* San Francisco: W. H. Freeman.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of threedimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences, 200*, 269–294.

Maunsell, J. H., & Newsome, W. T. (1987). Visual processing in monkey extrastriate cortex. *Annual Review of Neuroscience, 10*, 363–401.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29*, 317–322.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375–407.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.

Meuwese, J., Loon, A., Lamme, V., & Fahrenfort, J. (2014). The subjective experience of object recognition: Comparing metacognition for object detection and object categorization. *Attention, Perception, & Psychophysics, 76*, 1057–1078.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*, 275–292.

Mumford, D. (1992). On the computational architecture of the neocortex. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.

Nakamura, H., Gattass, R., Desimone, R., & Ungerleider, L. G. (1993). The modular organization of projections from area V1 and V2 to areas V4 and TEO in macaques. *Journal of Neuroscience, 13*, 3681–3691.

Navon, D., & Margalit, B. (1983). Allocation of attention according to informativeness in visual recognition. *The Quarterly Journal of Experimental Psychology Section A, 35*, 497–512.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* San Francisco: Freeman.

Norman, J. (2002). Two visual systems and two theories of perception: An attempt to reconcile the constructivist and ecological approaches. *Behavioral and Brain Sciences, 25*, 73–144.

Norman, D. A., & Bobrow, D. G. (1976). On the role of active memory processes in perception and cognition. In C. N. Cofer (Ed.). *The structure of human memory*. San Francisco: Freeman.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in Psychology, 4*, 124.

Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory–cognitive interaction. In D. A. Norman, & D. E. Rumelhart (Eds.). *Explorations in cognition* (pp. 279–307). Hillsdale, NJ: Erlbaum.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5*, 291–303.

Peyrin, C., Michel, C. M., Schwartz, S., Thut, G., Seghier, M., Landis, T., ... Vuilleumier, P. (2010). The neural substrates and timing of top-down processes during coarse-to-fine categorization of visual scenes: A combined fMRI and ERP study. *Journal of Cognitive Neuroscience, 22*, 2768–2780.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature, 343*, 263–266.

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51*, 1–41.

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 811–829.

Reinitz, M. T., Wright, E., & Loftus, G. R. (1989). Effects of semantic priming on visual encoding of pictures. *Journal of Experimental Psychology: General, 118*, 280–297.

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

Rock, I. (1983). *The logic of perception.* Cambridge: MIT Press.

Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., & Shevtsova, N. A. (1998). A model of attention-guided visual perception and recognition. *Vision Research, 38*, 2387–2400.

Schendan, H. E., & Maher, S. M. (2009). Object knowledge during entry-level categorization is activated and modified by implicit memory after 200 ms. *Neuroimage, 44*, 1423–1438.

Schendan, H. E., & Stern, C. E. (2008). Where vision meets memory: Prefrontal–posterior networks for visual object constancy during categorization and recognition. *Cerebral Cortex, 18*, 1695–1711.

Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., & Zetzsche, C. (2001). Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. *Journal of Electronic Imaging, 10*, 152–160.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory & Cognition, 23*, 681–696.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13, Whole No. 517).

Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*, 318–320.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews, 15*, 745–756.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1494–1505.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21*, 233–282.

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

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.

Tucker, T. R., & Fitzpatrick, D. (2003). Contributions of vertical and horizontal circuits to the response properties of neurons in primary visual cortex. *The Visual Neurosciences, 1*, 733–743.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition, 32*, 193–254.

Ullman, S. (1995). Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex, 1*, 1–11.

Vogels, R., & Orban, G. A. (1996). Coding of stimulus invariances by inferior temporal neurons. *Progress in Brain Research, 112*, 195–211.

von Helmholtz, H. (1867). *Treatise on physiological optics (J.P.C. Southall, Trans. (1962)), Vol. 3.* New York: Dover.

Wagar, B. M., & Dixon, M. J. (2005). Past experience influences object representation in working memory. *Brain and Cognition, 57*, 248–256 (Proceedings of TENNET 14).

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*, 167–194.

Walther, D. B., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.

Walther, D. B., & Koch, C. (2007). Attention in hierarchical models of object recognition. *Progress in Brain Research, 165*, 57–78.

Wyatte, D., Jilk, D., & O'Reilly, R. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology, 5*, 674.

Yarbus, A. L. (1967). *Eye movement and vision.* New York: Plenum Press.