Parallel visual search and rapid animal detection in natural scenes

Jan Drewes
Centre de Recherche Cerveau et Cognition (UMR 5549), CNRS, Université Paul Sabatier, Toulouse, France

Julia Trommershäuser
Center for Neural Science, New York University, New York, USA

Karl R. Gegenfurtner
Department of Psychology, Giessen University, Giessen, Germany

Human observers are capable of detecting animals within novel natural scenes with remarkable speed and accuracy. Recent studies found human response times to be as fast as 120 ms in a dual-presentation (2-AFC) setup (H. Kirchner & S. J. Thorpe, 2005). In most previous experiments, pairs of randomly chosen images were presented, frequently from very different contexts (e.g., a zebra in Africa vs. the New York Skyline). Here, we tested the effect of background size and contiguity on human performance by using a new, contiguous background image set. Individual images contained a single animal surrounded by a large, animal-free image area. The image could be positioned and cropped in such a manner that the animal could occur in one of eight evenly spaced positions on an imaginary circle (radius 10-deg visual angle). In the first (8-Choice) experiment, all eight positions were used, whereas in the second (2-Choice) and third (2-Image) experiments, the animals were only presented on the two positions to the left and right of the screen center. In the third experiment, additional rectangular frames were used to mimic the conditions of earlier studies. Average latencies on successful trials differed only slightly between conditions, indicating that the number of possible animal locations within the display does not affect decision latency. Detailed analysis of saccade targets revealed a preference toward both the head and the center of gravity of the target animal, affecting hit ratio, latency, and the number of saccades required to reach the target. These results illustrate that rapid animal detection operates scene-wide and is fast and efficient even when the animals are embedded in their natural backgrounds.

Keywords: visual search, rapid animal detection, natural scenes, natural context


Introduction

The human visual system is capable of analyzing newly presented scenes with remarkable speed and accuracy (Potter, 1975; Potter & Levy, 1969). In the classic manual go/no-go paradigm, where an observer is asked to detect the presence or absence of an object in a briefly presented natural scene, humans are able to correctly classify the presented stimulus in under 400 ms (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Thorpe, Fize, & Marlot, 1996). Event-related potentials show significant effects as early as 150 ms, a difference evidently related to an actual visual decision rather than basic stimulus properties (Johnson & Olshausen, 2003; VanRullen & Thorpe, 2001a, 2001b). This kind of rapid object detection has been found under various complex stimulus conditions, such as the detection of objects embedded in natural scenes, mostly animals, but also food (Fabre-Thorpe, Richard, & Thorpe, 1998), trees (Vogels, 1999a), and even artificial objects such as various vehicles (“means of transport,” containing cars, aircraft, boats, etc.; VanRullen & Thorpe, 2001a).

Generally, the processing related to object detection is attributed to the ventral pathway (Logothetis, Pauls, & Poggio, 1995; Tanaka, 1996; VanRullen, 2003). To a large part, the processing of visual information in the ventral pathway has been suggested to be implemented in a parallel fashion, allowing for the analysis of several scenes at the same time. This parallelism seems to have a limit though, as performance does decrease when the number of scenes depicted at the same time is increased (Rousselet, Fabre-Thorpe, & Thorpe, 2002). In a modified task, using eye movements instead of button presses, human observers have been shown to be capable of deciding which of two simultaneously presented natural scenes contains a target object in even less than 150 ms—even before the first discernible effect is visible in ERPs. On average, when using saccadic choice paradigms, human observers have been found to perform rapid animal detection with mean latencies around 200 ms, while maintaining accuracy ratings of 80% and up to 95% (Kirchner & Thorpe, 2005).

While the parallel nature of the ventral processing does not seem to be based simply on the possible existence of two lateralized independent ventral streams and is thus not
limited to a 2-fold parallelism, there appears to be a clear cost in processing an increasing number of scenes, both in terms of behavior and ERP data: in a manual response task, Rousselet et al. reported 74.7% mean accuracy and 493 ms mean response time for 2 images and 67.6% accuracy and 504 ms response time for 4 images (Rousselet, Thorpe, & Fabre-Thorpe, 2004). The exact nature of the mechanism used to recognize objects with such swiftness is still unknown; many attempts at reproducing human performance through computer algorithms have been undertaken. One of the more widely discussed theories was that the global amplitude spectrum of natural scenes may be used to perform this rapid, yet, accurate detection task (Johnson & Olshausen, 2003; Torralba & Oliva, 2003). While it has been shown that this kind of information representation does allow for classification performances approaching 80%, it has been shown that human subjects are able to perform the task even when this information is missing (Wichmann, Drewes, Rosas, & Gegenfurtner, 2010) and not when it is present and the phase spectrum is perturbed (Wichmann, Braun, & Gegenfurtner, 2006). In summary, despite of the existence of several intriguing theories, not even the precise nature of the data representation underlying this ultrarapid visual processing has become clear.

However, one might challenge the “naturalness” of the visual detection task used in the aforementioned behavioral experiments. Is detecting the presence of an object, when it is presented in a known location of the visual field—in the fovea even—the same as detecting an object when free-viewing ones surrounding? Thorpe et al. reported that the detection of animals in natural scenes works even when the images were shown in the far periphery of the visual field (up to 70.5° eccentricity; Thorpe, Gegenfurtner, Fabre-Thorpe, & Bülthoff, 2001). This experiment was certainly a step closer to a truly “natural” task, yet this experiment still used the same basic procedure and stimulus design that all the previous experiments shared: even during the single/multiple stimulus display, there was a neutral background surrounding the images shown on the screen, theoretically reducing the portion of the visual field that needed to be analyzed by the subjects by a huge portion. Furthermore, in the experiments by Thorpe et al. where multiple scenes were shown simultaneously, these individual scenes were usually completely unrelated to each other (e.g., the New York skyline vs. a zebra in Africa) and physically well separated.

Most of these experiments were performed using professional photographs (from the Corel Stock Photo Library). This selection of the stimulus material might have affected stimulus processing. Context congruence has generally been shown to have a significant effect on classification accuracy and speed (Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008; Rieger, Koechy, Schalk, Grueschow, & Heinze, 2008).

These results suggest that unrelated sceneries might have a strong target/background congruence difference, allowing for a “holistic” kind of object detection, where the absence or presence of an object might be judged simply by the general gist of a scene rather than the perception of the actual object. This kind of classification has been modeled to a relatively high degree of success (e.g., Oliva & Torralba, 2001, 2006; Torralba & Oliva, 2003), but although the performance of such a classifier resembles typical human behavior in some cases (Greene & Oliva, 2009), it cannot explain human performance (Wichmann et al., 2010) in the animal detection task.

Even more worrisome, in most experiments, the location of the upcoming stimulus image was either known to the subject or had a low degree of uncertainty (except for Thorpe et al., 2001), and the target object was always depicted in or very near to the center of the image. This did not require subjects to perform an actual search of the target object but constrained the detection task to a limited number of clearly separated candidate locations to be tested. In the classic 2AFC paradigm, it is therefore possible for a subject to employ a proxy strategy: instead of answering whether the object appeared on “the left or the right image,” the subjects can focus their attention on just one of the candidate locations, answering whether the object appeared “on the left image or not.” While this proxy strategy might lead to equivalent results in the laboratory, it would certainly not be applicable in real-life free viewing.

In addition, the parallelism of the visual processing might be better equipped to analyze one contiguous scene, possibly spanning the entire visual field, rather than several independent scenes covering only a fraction of the observer’s view.

While previous studies have made a strong point about the general capability of the human visual system to perform ultrarapid object detection with high efficiency, the applicability of these results to real-life viewing conditions has not been demonstrated.

In this study, we used a new set of stimuli in a modified saccadic choice paradigm to analyze human performance in both the classic 2AFC animal detection task and also in a more natural setting, offering a high degree of uncertainty in target location with large, contiguous stimuli. This demanded that our subjects perform a genuine visual search in order to find their target, eliminating the possible use of simple proxy strategies.

**Methods**

**Subjects**

Ten observers participated in all 3 experiments (3 males, 7 females, aged 23–33 years). All participants had normal or corrected-to-normal vision and were naive to the purpose of the experiment. The participants were students...
recruited from the University of Giessen and paid hourly for their participation.

Setup

Eye movements were recorded using an EyeLink II camera-based system, manufactured by SR Research, Canada. The recording frequency was set to 250 Hz, using combined pupil and cornea reflex measurement. Subjects were seated at 45 cm away from the CRT screen, the active display of which extended approximately 48.3° horizontally and 38.6° vertically, at a resolution of 1280 × 1024 pixels. The refresh rate of the screen was set to 100 Hz, driven by a PC computer system. Stimulus presentation was realized with a proprietary software written in C. The experimental chamber had all major surfaces covered in black and was otherwise dimly lit; the exact amount of lighting was not calibrated, although care was taken to achieve the same level of lighting throughout all experimental sessions. At the beginning of each session, subjects were calibrated using the default 9-point calibration algorithm suggested by the manufacturer.

Experiment 1: 8-Choice

The participants in Experiment 1 performed a combined visual search and rapid animal detection task. We monitored the subject’s eye movements through all phases of the search task. Most of the recent studies involving rapid animal detection used images drawn from the Corel Stock Photo Libraries (CSPL), which offer professional quality but comparatively low-resolution images. The center objects in the CSPL frequently cover a large portion of the available image area, leaving relatively little target-free surround. In order to create stimuli with a large, contiguous, yet target-free surround, larger (and higher resolution) images were needed in which the center object, if any was present, did not cover a majority of the image. Such images can be found in the newly collected All Natural Images Database (ANID) (2011). From the several thousand images in the ANID, we selected those that offered either one single animal or a small and compact group of animals with a large, animal-free surround on all sides of the animal. The image was then cropped and downsized (when necessary) to a square shape of approximately 1600-pixel side length, such that the animal was positioned in the center. This created stimuli with a contiguous background that was considerably larger than in previous studies. These stimuli maintain animal/background integrity and thereby avoid any kind of “cut out” or “cut and paste” effect, which can affect human response characteristics (Joubert et al., 2008). Of the images in the ANID, 294 were found to be usable for the above procedure (see Figure 1 for a selection of samples).

Finally, the scene depicted in any given trial was shifted and cropped in such a way that the animal appeared with equal probability in either of 8 possible locations, evenly spaced on an imaginary circle around the center of the screen (left, up, right, down, and on the four diagonals in between, see Figure 2). The radius of the circle was 10.5°, and the screen was superimposed with a centered circular aperture of 38° diameter, such that the extent of the image from screen center was the same in all directions. A set of sample stimuli can be seen in Figure 3 (“8-Choice”). During the experiment, every animal was shown in each of the 8 positions, totaling 2352 trials, partitioned in 3 sessions of 784 trials each. The order of trials was randomized to eliminate predictability of stimulus image or animal position. This also allowed us to include subjects in our analysis that did not complete all 3 sessions. Four subjects were tested in 1 session, 1 subject was tested in 2 sessions, and the remaining 5 subjects were tested in all 3 sessions.
**Paradigm**

Each trial started with the display of a fixation dot (about 0.1° of visual angle) in the center of the otherwise uniformly gray screen. Upon a button press by the subject, the fixation dot changed into a small circle (about 0.15° of visual angle) to signal the start of the trial. After a random interval of 300–800 ms, the fixation circle disappeared, followed by a gap period of 200 ms. After the gap period, the main stimulus image was displayed. While the stimulus display remained unchanged, subjects were allowed up to 5 s to direct their gaze to within 0.6° of the shape of the animal; at this distance, at least a small portion of the animal would have been projected inside their fovea. If the subject succeeded to move his/her gaze onto or to within the required proximity of the animal before the 5 s expired, the trial was considered successful, and the display then vanished after an additional delay of 500 ms, ending the trial. If the 5 s expired without the gaze of the subject reaching the animal, the trial was considered unsuccessful, and the stimulus display vanished, ending the trial. Before each trial, a drift correction was performed to maintain a high level of eye-tracking accuracy; after each block of 200 trials, subjects were permitted to take a break if they so desired. Prior to the experiment, subjects were instructed to direct their gaze onto the animal as quickly and accurately as possible; they were informed that the animal could appear “anywhere on the screen, except for the center region,” without explicit declaration of the 8 possible target locations. Subjects were not instructed to look at any particular part of the animal.

**Experiment 2: 2-Choice**

Experiment 2 differed from Experiment 1 solely in that only the horizontal target positions (left and right) were used; subjects were instructed that the animal could appear “either to the left or to the right.” This reduction in position uncertainty made the experiment similar to the classic 2AFC paradigm as originally used by Thorpe et al., while still maintaining the large, contiguous background (see Figure 3, “2-Choice”). The experiment was conducted in one session, containing 584 trials.

**Experiment 3: 2-Image**

In this experiment, the circular mask used in Experiments 1 and 2 was replaced by two squares, letting the underlying stimulus appear as two individual images (see Figure 3, “2-Image”). The size of each square was 17 × 17 degree visual field, each placed at 10.5-degree horizontal eccentricity (measured from the center of the square to the center of the screen). Through this, Experiment 3 resembled the visual appearance of the classic...
2AFC paradigm, while using the same contiguous underlying images as Experiment 2. With the exception of the shape of the mask, the stimuli and number of trials were identical to Experiment 2.

Analysis

Human performance

We recorded decision latency and hit ratio by measuring the latency (i.e., saccadic reaction time) and destination of the first saccade in each trial. During the post-hoc analysis, all trials were marked invalid in which the subject had drifted by more than 2° during fixation, or in which the first saccade was shorter than 2.5°. Saccades faster than 100 ms were considered as errors, possibly driven by the sudden appearance of the stimulus rather than stimulus content, and were also marked invalid. Invalid trials were ignored during further analysis. For “hit” detection, in previous experiments, it was considered sufficient to test whether the destination of the first saccade was located anywhere inside the correct image (the one showing the target stimulus). For the 8-Choice design, it would in theory suffice to radially divide the displayed circle into 8 evenly sized sectors; however, the shape of many of the animals in our selection would extend outside of the limits of one such sector, allowing for saccades that actually hit the animal, or went in a correct direction, but ended outside the proper sector. These would be counted as misses, negatively biasing hit detection. At the same time, counting whether the first saccade actually landed directly on the shape of the animal proved to be unreliable, since many subjects produced hypometric saccades; subjects would therefore frequently need more than one saccade to actually arrive on the shape of the animal, even though the close proximity of the first saccade endpoint to the animal clearly suggested that they had already detected their target. A trial was counted as “hit” if the first saccade went into precisely that sector of the image that just covered the extent of the shape of the animal (see Figure 4); consequently, every saccade that went into a direction intersecting with any part of the animal was counted as a hit, regardless of the actual distance to the animal.

As can be seen from Figure 4, this method can create “hit” sectors of different sizes for different display positions of the same animal, possibly favoring certain positions on the imaginary circle. To estimate the direction of these sectors, we divided the circle into 8 evenly sized sectors, and marked the first and last part of each sector as hit zones. Any saccade that landed in a hit zone was counted as a hit, regardless of the actual distance to the animal.
and magnitude of possible biases, we computed the mean distribution of animal shapes by aligning all animal shapes on their individual centers of gravity, and then superimposing them in a single plot (Figure 5). Overall, the used animal shapes were quite evenly distributed, on average just slightly wider than tall. Consequently, across all images, the angular size of the eight target sectors was approximately evenly distributed, with possibly only a slight (symmetric) bias favoring vertical directions.

Latency was determined as the time from stimulus onset to the onset of the first saccade, as reported by the EyeLink systems built-in saccade detection (thresholds: speed of 22 deg/s and acceleration of 3800 deg/s²).

Spectral image properties

As was previously shown, the global amplitude spectrum does not serve as a primary cue in human animal detection (Wichmann et al., 2010). For completeness reasons, and since we are introducing a new set of images in this work, we nonetheless decided to compare the average amplitude spectra of the animal and non-animal rectangles of our stimuli as used in Experiment 3. The average amplitude spectra were extremely similar; the mean difference between both the average animal and the average non-animal amplitude spectra was just less than 10% of one standard deviation, with a maximum of 73%. With less than 300 images in each class, these very slight differences are easily explained by noise alone, indicating that no viable cue of animal presence could have been contained in the global spectral information.

Saliency model

The performance of our subjects may be partially related to the distribution of low-level saliency in our stimuli. In an attempt to explore the possible extent of this relation and the possible contribution of low-level saliency to the performance of our subjects, we employed the “Saliency Toolbox” provided by Walther and Koch (2006). Saliency maps were computed on the full target images. These saliency maps were then translated and cropped in the same manner as our original stimuli (see Experiment 1). After the circular mask was applied, the winner-take-all algorithm was used to compute simulated saccades using the default parameters suggested by the toolbox. To prevent simple return saccades, the inhibition-of-return (IOR) feature was utilized with an inhibition area of circular shape and 2.5-degree equivalent radius. Exploratory testing showed that, on many images, no more than 5 saccades were produced, because all saliency in the image had been excluded by the IOR, while smaller IOR areas frequently resulted in multiple saccades close to the same visual feature of the processed image. The number of incremental steps of the neural network was therefore adapted to simulate a maximum of 5-s viewing time or 5 saccades, whichever happened first.
Results

Experiment 1

Due to the 8 possible target positions, among which the target animals were uniformly distributed, chance performance was at 12.5% hit ratio; however, all 10 subjects exceeded chance performance, yielding, on average, 80.1% hit ratio at a mean latency of 207.5 ms. Table 1 summarizes the performance of all subjects measured in Experiment 1.

Directional analysis: Bias toward upper hemisphere

A comparison of detection performance across target locations showed slightly better performance for animals presented in the upper hemisphere (see Figure 6A). Most efficient detection performance occurred for stimuli presented at the “up” position (86.6% correct, pooled across all 10 subjects) and least efficient detection for the “down” position (72.1% correct). In general, when pooling the “upper left,” “up,” and “upper right” positions to represent the upper visual field, and the “lower left,” “down,” and “lower right” positions as the lower visual field, the hit ratio for targets in the upper visual field is, on average, higher (82.0% correct) than for targets in the lower visual field (75.8% correct). This difference is significant (paired t-test, \(df = 9, p = 0.002\)). At least a part of this difference may be due to the distribution of low-level saliency across the scenes (see below and Figure 6B). Overall, latencies were fast (207.5 ms, pooled across all 10 subjects) and slightly slower for animals placed in the lower visual hemisphere (210 ms) compared to the upper hemisphere (204 ms). This difference is also significant (paired t-test, \(df = 9, p = 0.023\)). No further significant trends in directionality were found; a preference for the cardinal directions as suggested by previous work on natural scenes (Einhauser et al., 2007) was either not present in the data or hidden by the difference between the upper and lower visual fields.

Experiment 2

Chance performance was 50%; however, all subjects performed better than chance, scoring, on average, 90.3% correct, at a mean latency of 200.4 ms. The performance of all subjects for Experiment 2 is shown in Table 2.

Experiment 3

As in Experiment 2, chance performance was 50% correct. Once again, all subjects performed above chance, yielding, on average, 88.6% correct, at a mean latency of 198.1 ms. The performance of all subjects for Experiment 3 is shown in Table 3.

Comparison between experiments

Hit ratio and latency: No significant difference after normalization

Of the three experiments run in this study, Experiment 1 offered the subjects with 8 alternative target locations, while the other two only offered 2 alternatives, located on the horizontal axis. While a direct comparison of the detection performance reveals approximately 10% better performance on both of the 2-alternative experiments (10.5% on Experiment 2; 8.2% on Experiment 3), such a direct comparison is not adequate. Where both 2-alternative experiments come with an implicit native chance performance of 50%, the 8-alternative experiment only offers 12.5%. Furthermore, mean classification performance in the horizontal direction was, on average, higher than in the vertical direction (Figure 6A). For better comparability, we normalized the results of all 3 experiments by rescaling the above-chance range to a common virtual range of 0–100%. Our scaling is similar to the guessing corrections applied in classical psychophysics (Gescheider, 1997). While other means of normalization may be considered, the chosen technique compensates for the different chance performance levels while maintaining an intuitively understandable scale; instead of absolute, hit

<table>
<thead>
<tr>
<th>Subject</th>
<th>(N) sessions</th>
<th>(N) trials</th>
<th>(N) valid trials</th>
<th>Hit ratio (%)</th>
<th>Hit latency (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>784</td>
<td>669</td>
<td>80.6</td>
<td>225.4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2352</td>
<td>1944</td>
<td>79.9</td>
<td>172.1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1568</td>
<td>870</td>
<td>78.1</td>
<td>212.7</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1568</td>
<td>1714</td>
<td>77.0</td>
<td>200.4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>784</td>
<td>590</td>
<td>81.6</td>
<td>206.3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1568</td>
<td>1541</td>
<td>80.2</td>
<td>199.9</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1568</td>
<td>2196</td>
<td>83.6</td>
<td>197.2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>784</td>
<td>637</td>
<td>83.5</td>
<td>216.5</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>2352</td>
<td>2060</td>
<td>84.7</td>
<td>220.0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>784</td>
<td>681</td>
<td>71.9</td>
<td>224.1</td>
</tr>
<tr>
<td>Mean</td>
<td>2.1</td>
<td>1646</td>
<td>1290.2</td>
<td>80.1</td>
<td>207.5</td>
</tr>
</tbody>
</table>

Table 1. Experiment 1, general animal detection performance.
ratios are expressed relative to the theoretical performance minimum and maximum. Additionally, we analyzed the trials in which the target was presented at either of the horizontal positions of Experiment 1; the comparison between experiments was based only on the horizontal target positions. Latencies were not normalized but also analyzed separately for the horizontal target positions. The general results before and after normalization are shown in Table 4.

After normalization, the mean horizontal hit ratio of the 10 subjects in Experiment 1 (8-Choice) and Experiment 2 (2-Choice) is almost identical; the small difference in latency (6.3 ms) is statistically significant (paired t-test, $df = 9, p = 0.01$); however, it should be considered that a difference of 6.3 ms is less than 2 cycles of the eye-tracking system used (250-Hz tracking frequency, 4-ms cycle time). The hit ratio in Experiment 3 (2-Image) is 4.7% less than in Experiment 1, which is marginally
significantly slower for the 8-Choice experiment despite the increased number of targets (paired \(t\)-test, \(df = 9, p = 0.048\)). Generally, it should be noted that individual sessions corresponding to the different experiments were recorded on different days, allowing for some day-to-day variance within subjects.

**Timing: Increased position uncertainty does not cause increased latency**

As a means of determining the fastest systematically correct responses our subjects were capable of, we accumulated the hit and miss latencies of our subjects in bins of 4-ms width, corresponding with the sampling frequency of our eye-tracking system (250 Hz). Subsequently, we computed \(X^2\) tests (\(df = 1\)) on every interval. The first above-chance bin was then determined as the start of the first sequence of at least 3 significant test results. Sample histograms for our fastest and slowest subjects are shown in Figure 7; the earliest significant time intervals estimated for correct detection in all 3 experiments are shown in Table 5.

On average, the first correct responses occur at 160 ms in Experiment 1 (8-Choice) and around 164 ms and 162 ms in Experiments 2 and 3 (2-Choice). This difference was not significant (repeated measures ANOVA, \(df = 2, F = 0.873, p = 0.43\)). Note that saccade latency is not slower for the 8-Choice experiment despite the increased number of alternative target locations (see Discussion section).

### Target zones

Any target of a complexity comparable to that of animals will have multiple features that may serve as cues for the visual system. The precise nature of these features is still largely unknown. While the current experimental design might not be suitable for a detailed analysis of the importance of an arbitrary number of individual features, the spatial distribution of saccades on the animals might still provide some hints as to which areas of the animal are of interest to the visual system.

#### A visible head improves detection performance

As our targets were, on average, displayed at 10.5° eccentricity, the center of gravity of the shape of the animal is one of the more likely candidates for a primary saccade target: it can easily be computed even with the reduced resolution of the retinal periphery. As a second possible target, the head of the animal was included in our analysis, as it is the part of the animal that will usually convey the most information about the animal’s focus of attention and intentions (hunt, flee, whether it has noticed the observer, etc.). All perspectives of the head were counted as a “visible” head, not only frontal or profile views. Of our 294 stimulus images, 233 showed the head of the animal, while 61 did not. Hit ratio was generally higher and response times were faster for animals with visible heads than for animals without visible heads. Table 6 shows the hit ratio and Table 7 shows the hit latency for Experiment 1, averaged over all 10 subjects. The visibility of the head significantly affected performance in 5 out of 6 measures; a comparison of the horizontal trials of all three experiments and the results of the corresponding paired \(t\)-tests (\(df = 9\)) can be seen in Table 8.

#### Head and center of gravity attract saccades

We further analyzed the trials in which the subject’s gaze eventually reached the animal. We compared performance depending on the number of saccades needed to hit the animal. Figure 8 shows the saccade distribution for the set of stimuli shown in Figure 1. Figure 9 illustrates the distribution of the number of saccades required to reach the target animal. In 68% of all trials, the shape of the animal was reached with the first saccade, and in approximately 20% of all trials, a second saccade was needed. In 72% of the trials in which the head of the animal was visible, the shape of the animal was reached with the first saccade; this was the case in 54% of the trials with those animals in which the head was not visible.

When aligning all animal shapes along their centers of gravity, the distribution of animal area is approximately spherical (see Figure 5A). The heads our animals are...
almost exclusively located above the center of gravity of the animals (Figure 5B)—an anatomically unsurprising finding. The distributions of animal and head-only area therefore have different centers, a fact that we exploited for more analysis, by excluding all animals without visible heads. In order to normalize landing zones across the various animal shapes, the following transformation was derived individually for all images:

1. Translation: All images were centered on the center of gravity of the depicted animal.
2. Rotation: All images were rotated around their individual center of gravity, such that the center of the head area was then located on the horizontal axis, exactly to the right of the center of gravity.
3. Symmetric scaling: All images were symmetrically scaled with respect to the center of the scaling on the center of gravity, in such a way that the center of the head would always come to rest at the same (virtual) distance to the center of gravity of the animal.

We finally compared the distribution of all saccade endpoints across all individual images, by plotting them into a common virtual scale. Every saccade endpoint was represented by a Gaussian blob of initial diameter, which was also scaled according to the individual transformation of the respective image. A visualization of the result can be seen in Figure 10. To obtain a numerical representation of the distribution of saccades between the center of gravity and the head, square regions of identical size were centered on both head area and center of gravity: Table 9 specifies the percentage of saccades that landed in either the center of gravity (COG) or the head region.

A comparison of saccades that were counted as either head- or COG-directed revealed that the distribution of saccade endpoints shifted, depending on the number of saccades required to hit the target: in those trials in which the first saccade was a direct hit, about 60% of all saccades in the two squares were closer to the COG than the head; in those trials where the second saccade hit the target, about 47% were directed toward the COG, and finally, in all trials that required 3 or more saccades, about 60% of the saccade endpoints were closer to the head of the target animal. The trend and individual subject data can be seen in Figure 11.

Additionally, we analyzed the spatial distribution of the first saccades depending on saccade latency. Similar to above, saccades were projected onto an imaginary line passing through the COG and the head. Saccades that ended directly on the animal’s head were assigned a value of 1, those that landed exactly on the center of gravity were assigned a value of 0, and those exactly in the middle between our two points of interest were assigned a weight of 0.5. To be able to focus on saccades that were indeed targeted toward either the head or the COG, saccades with a distance to either the COG or the head that was more than four times the distance between head and COG were eliminated, as well as saccades with resulting weights of less than −1 or more than 2. Figure 12 shows the distribution of the remaining latencies \(N = 2645\) by assigned saccade value. By fitting a line to the values, we found no significant slope (least-squares fit, slope of −0.013 relative to z-scored latencies on the interval \([-1, 2]\)). We continued to pool the latencies separately for each subject in 5 evenly spaced intervals (1: \([-0.75, 0.25]\), 2: \([-0.25, 0.25]\), 3: \([0.25, 0.75]\), 4: \([0.75, 1.25]\), 5: \([1.25, 1.75]\)). The intervals were arranged in such a way that both COG and the head were at the center of one interval. This would have allowed us to recognize possible peaks or minima near our points of interest. A statistical test showed no significant difference between intervals (repeated measures ANOVA, \(df = 4\), \(F = 0.814\), \(p = 0.52\)).

**Low-level saliency**

In an attempt to investigate the vertical asymmetry in hit ratio, and to analyze the possible contribution of low-level saliency to our subjects’ task performance in general, we referenced the location of the saliency-driven saccades as predicted by the winner-take-all algorithm with the location of the animals and the target sectors of the individual images. Sample images, corresponding saliency maps, and generated saccades can be seen in Figure 13.

**Comparatively low saliency-based hit ratio**

Across all target positions, the first predicted saccade ended on the animal shape in 24.1% and within the hit sector in 41.4% of all simulated trials (see Table 10). The target position-dependent performance is illustrated in Figure 6.
Figure 7. Latency histograms of the fastest (S2) and slowest (S9) subjects. The vertical black lines mark the first/fastest bin of significantly above-chance performance.
Direct hits on the first simulated saccade occurred almost 3 times as often when the head of the animal was visible, reaching 27.8% for visible heads vs. 10.0% for invisible heads. On the second saccade, this difference was reduced and inverted for the 3rd and 4th saccades. The first saccades ending inside the hit sector (see Figure 4) showed a similar tendency, reaching 45.4% hit ratio when the animal’s head was visible and 25.8% when the head was not visible. While certainly better than chance performance, these hit ratios are far less than those achieved by our human subjects.

High saliency density on the target animals

Relative to all the saliency contained in any one scene, an average of 22.5% of all saliency falls within the shape of the target animals, which cover only about 1% of the visible image area (see Table 11). In those cases where the head of the animal is visible, on average, 2.5% of the saliency falls within the shape of the head, which, on average, covers a tiny 0.13% of the visible image area. This leads to an average animal saliency density (saliency per image surface area covered by the animal) that appears to be phenomenally high; however, there may be other small objects in the remainder of the scene that possess similarly salient features (see Figure 13). While these objects may easily act as distractors to the saccade prediction algorithm, due to the size of the remaining scene, these salient objects would not significantly increase the average saliency density of the remaining scene. This can explain why the saliency model does not perform as well as the human subjects. Interestingly, the average saliency density on the animal heads is slightly less than that on the entire animals; the saliency density alone cannot, therefore, explain the improved performance on images with visible heads apparent with both humans and the saliency model itself.

Discussion

Summary

Overall, all our subjects performed at high levels of speed and accuracy in all three experiments. We found that:

1. Rapid animal detection is still possible in a large, contiguous scenery instead of just isolated image frames and is even slightly facilitated.
2. Rapid animal detection is independent of the number of possible target positions and must, therefore, be implemented in parallel.
3. Subjects cannot only detect but reliably locate the targets even in the presence of multiple candidate locations.
4. At least two main target zones tend to attract the saccades of human observers, which are the center of gravity and the head of the animal.
5. The presence of a visible head significantly improves detection performance, although all subjects still performed well when no head was visible.

Stimuli and design

In any study examining the processing of natural scenes in humans as well as other mammals, the question of how “natural” the scenes presented to the observer really are is always difficult to answer, as there exists no canonical measure (or scale) of “naturalness.” Frequently, any photography of a real-world scene is considered a natural scene; sometimes, specifically in object detection, object/background consistency are also considered: a TV set on a rooftop will not usually be considered to be consistent, yet since it is possible to take such a picture, the said picture could technically be considered to be a natural scene. A more naturalistic view would be to allow a scene to contain only naturally occurring objects, excluding man-made objects and structures. In our everyday life,
however, man-made objects are very frequent, which is why we chose to allow some artificial items like huts and fences to occur in our stimuli. What all conventional photography is missing, however, is the 3rd dimension. While the technology to reproduce both static and dynamic scenes in 3D has been used in various psychological studies, high-quality image/scene databases usable for experimental designs similar to the one presented in this work have yet to be created. Therefore, our study was ultimately limited to the 2D computer displays and projection systems commonly found in research setups. In this study, several aspects of the naturalness of our stimuli were carefully considered. All photographs used in stimulus creation were taken in wildlife parks and zoos where the animal enclosures were at least superficially modeled to the individual animal’s natural surroundings. While vegetation and climate settings might differ from the animal’s native habitats, with the exception of a few barns and fences, the entire scene shown in the photographs had a very authentic appearance and did generally not include any dominating man-made structures. The actual stimuli used in our study are unique also because of their size and background contiguity. The high-resolution images we took from the All Natural Images Database allowed the image circle displayed in Experiments 1 and 2 to be of a rather large size ($38^\circ$ visual angle). The presentation of 2 visually distinct images, such as the New York skyline in one location and some African wilderness in the second location, can give low-level cues about the probability of the presence of certain objects, including animals, in one or both of the stimulus images. We avoided exposing our subjects to these low-level cues by maintaining image congruity in all conditions. Due to the size of the source images, there was enough margin around the target animal to shift the entire scene, such that the animal could be found in various locations in the image, rather than only in the image center, without breaking image contiguity. This allowed target and no-target locations to be taken from one and the same scene, without a contextual contrast as it can occur in simultaneous presentations of 2 individual images. The circular shape of our stimuli ensured that the image extended evenly from the initial point of fixation, avoiding possible artifacts induced by the rectangular corners of square-shaped images. However, shifting the image also caused the visible part of the stimulus image to vary with the target position. This may have contributed to the difference in classification performance in the vertical hemispheres (see below). A potential pitfall in a 2-alternative forced-choice experiment is that the observer might focus all attention on just 1 of the 2 locations, thereby answering the question of whether the target appeared at that one location or not, rather than whether it appeared at the one location or the other. While this strategy is theoretically a valid solution to the task (identifying 1 out of 2 locations), it, of course, evades solving the originally intended problem (simultaneous analysis of 2 locations). More importantly, such a simplified proxy strategy would have no relevance in real-life free viewing. We excluded the use of such simple proxy strategies by increasing the number of alternative target locations (from 2 in Experiments 2 and 3 to 8 in Experiment 1). Had our subjects still focused their attention on just one of the 8 locations, they would have suffered a severe impairment to their detection performance; this was not the case (see below). We therefore believe that our stimuli were in fact as natural as they can be in the physical setup used to run our experiments.

**Performance**

While our subjects performed reasonably symmetrically for target animals presented on the horizontal axis, there

<table>
<thead>
<tr>
<th>Experimental measurement</th>
<th>Animals with visible heads</th>
<th>Animals without visible heads</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1 (8-Choice, horizontal only)</td>
<td>Hit ratio 84.5%</td>
<td>79.9%</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Hit latency</td>
<td>205.1 ms</td>
<td>213.4 ms</td>
</tr>
<tr>
<td>Experiment 2 (2-Choice)</td>
<td>Hit ratio 90.8%</td>
<td>88.4%</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>Hit latency</td>
<td>198.9 ms</td>
<td>207.6 ms</td>
</tr>
<tr>
<td>Experiment 3 (2-Image)</td>
<td>Hit ratio 89.2%</td>
<td>86.5%</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>Hit latency</td>
<td>197.6 ms</td>
<td>204.2 ms</td>
</tr>
</tbody>
</table>

Table 8. All experiments: Comparison of effect of visible head.
Figure 8. Saccade endpoint distributions on sample animals (see Figure 1). Blue dots denote the endpoints (landing coordinates) of saccades that did not hit the animal; red dots denote hits. The black border around the animal shapes illustrates a margin of 0.5 degree.
was a clear bias in detection accuracy toward the upper visual hemisphere. This vertical asymmetry may have been caused by general differences in saliency and clutter between the upper and lower halves of the photographs used to create our stimuli: since we shifted the entire scene rather than moving the target object within the scene, the visible portion of the upper half of the stimulus image was larger when the target animal was presented in the lower visual hemisphere, and smaller when the animal was presented in the upper hemisphere. However, as our comparison between the 3 different experiments is based on the horizontal stimulus positions only, this did not influence the quality or validity of our main results. In general, the differences in the performance of our subjects between the 3 different experiments were surprisingly small; the average latencies were comparable to those shown in previous studies (e.g., Thorpe et al., 1996). The approximate 10% difference in raw hit ratio between Experiment 1 and Experiments 2 and 3 was alleviated by the normalization of the data to identical scales. When looking at the horizontal target positions only, we did in fact find a statistically significant difference between Experiment 1 (8-Way) and Experiment 3 (2-Image) for both latency and hit ratio. Note that the average hit ratio was actually highest on Experiment 1 (8-Way, 81.3%), even though the number of possible target locations was 4 times as high as in Experiment 3 (2-Image, 76.6%). The difference in latency between the 8-Way experiment and the 2-Way experiments was also significant, with mostly faster latencies in Experiments 2 and 3 (2-Way/2-Image). Because of this, a speed/accuracy trade-off might explain some of the difference in hit ratio between Experiments 1 and 3—the difference in latency is about 4.3% relative to the hit ratio.
(8.6 ms absolute), while the difference in hit ratio is 6.1% relative (4.7% absolute). A further comparison of the different experiments makes such a trade-off rather unlikely—the comparison between Experiments 1 (8-Way) and Experiment 2 (2-Way) shows no significant difference in hit ratio, although there is a significant difference in latency (6.3 ms), similar to the difference between Experiment 1 (8-Way) and Experiment 3 (2-Image). This 2-step difference between experiments (only latency differs between Experiment 1 and Experiment 2, only hit ratio differs between Experiment 2 and Experiment 3) suggests that the difference in hit ratio between Experiments 1 and 2 and Experiment 3 is caused primarily by the change in stimulus display (large, contiguous circle vs. 2 separate squares), while the difference in latency is

<table>
<thead>
<tr>
<th>Subject</th>
<th>COG (%)</th>
<th>Head (%)</th>
<th>Rel. head (%)</th>
<th>COG (%)</th>
<th>Head (%)</th>
<th>Rel. head (%)</th>
<th>COG (%)</th>
<th>Head (%)</th>
<th>Rel. head (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.3</td>
<td>16.2</td>
<td>42.1</td>
<td>33.6</td>
<td>19.3</td>
<td>36.5</td>
<td>19.2</td>
<td>20.7</td>
<td>51.8</td>
</tr>
<tr>
<td>2</td>
<td>21.5</td>
<td>16.5</td>
<td>43.5</td>
<td>27.4</td>
<td>22.9</td>
<td>45.6</td>
<td>13.9</td>
<td>19.0</td>
<td>57.6</td>
</tr>
<tr>
<td>3</td>
<td>36.7</td>
<td>17.6</td>
<td>32.5</td>
<td>18.1</td>
<td>22.4</td>
<td>55.3</td>
<td>5.4</td>
<td>13.2</td>
<td>70.9</td>
</tr>
<tr>
<td>4</td>
<td>23.8</td>
<td>23.4</td>
<td>49.6</td>
<td>16.2</td>
<td>32.6</td>
<td>66.9</td>
<td>8.2</td>
<td>21.6</td>
<td>72.5</td>
</tr>
<tr>
<td>5</td>
<td>34.0</td>
<td>17.0</td>
<td>33.3</td>
<td>20.7</td>
<td>21.4</td>
<td>50.8</td>
<td>13.5</td>
<td>12.1</td>
<td>47.3</td>
</tr>
<tr>
<td>6</td>
<td>30.3</td>
<td>20.2</td>
<td>33.3</td>
<td>17.5</td>
<td>25.6</td>
<td>59.4</td>
<td>7.2</td>
<td>11.9</td>
<td>62.3</td>
</tr>
<tr>
<td>7</td>
<td>41.0</td>
<td>19.1</td>
<td>31.8</td>
<td>25.5</td>
<td>22.0</td>
<td>46.3</td>
<td>15.0</td>
<td>20.8</td>
<td>58.0</td>
</tr>
<tr>
<td>8</td>
<td>27.8</td>
<td>24.8</td>
<td>47.1</td>
<td>28.0</td>
<td>27.1</td>
<td>49.1</td>
<td>19.4</td>
<td>18.2</td>
<td>51.7</td>
</tr>
<tr>
<td>9</td>
<td>35.7</td>
<td>22.0</td>
<td>38.1</td>
<td>22.7</td>
<td>25.3</td>
<td>52.6</td>
<td>12.7</td>
<td>16.1</td>
<td>55.9</td>
</tr>
<tr>
<td>10</td>
<td>29.4</td>
<td>24.3</td>
<td>45.3</td>
<td>16.6</td>
<td>26.0</td>
<td>60.9</td>
<td>9.0</td>
<td>23.1</td>
<td>72.0</td>
</tr>
<tr>
<td>Mean</td>
<td>30.9</td>
<td>20.1</td>
<td>40.3</td>
<td>22.3</td>
<td>24.8</td>
<td>52.3</td>
<td>10.3</td>
<td>16.7</td>
<td>59.7</td>
</tr>
</tbody>
</table>

Table 9. Experiment 1: Distribution of saccades. “Relative” values were computed with only the “Head” and “COG” zones, ignoring the surround.

![Figure 11](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933481/) Relative distribution of saccade endpoints between COG and head, depending on the number of saccades required to reach the animal.

![Figure 12](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933481/) Latency distribution of the first saccade, relative to the distance to COG and head. The x-axis shows the distance from both COG and head, projected to a line going through both COG and head. The red line represents the least-squares line fit to latency values; the slope is 0.01 relative to z-scored latency values within the interval [-1, 2].
caused primarily by the increase in target location alternatives (2 vs. 8). Based on these results, we conclude that while the actual stimulus area to be analyzed was larger in Experiment 2 compared to Experiment 3, classification in terms of hit ratio was actually facilitated by the contiguous stimulus display. The difference in response latency between Experiments 1 and 2 was only about 6.3 ms (around 3.1%) even though the number of alternative target locations changed from 2 to 8. This may be a surprising result, since, generally, Hick’s law (Hick, 1952; Hyman, 1953) predicts decision latencies to increase with the logarithm of the number of alternative choices. However, saccadic responses do not always obey Hick’s law (Kveraga, Boucher, & Hughes, 2002; Lee, Keller, & Heinen, 2005), and latencies can sometimes even exhibit an anti-Hicks effect, which has most recently been shown by Lawrence, St. John, Abrams, and Snyder (2008). The experiment by Kveraga et al. was similar to our Experiment 1 with respect to several aspects: they also presented 8 targets on a circle of the same diameter as ours (10.5\degree), although they used simple targets (disks of 1\(^{\circ}\) diameter) instead of natural scenes and objects. They did not find the increase in latency predicted by Hick’s law; in fact, they found no significant difference in latency. While we found a statistically significant increase in latency, this difference was clearly not consistent with the log\(_2\) increase predicted by Hick’s law (3.1% real increase vs. 200% prediction). The similarity between paradigms makes it very likely that we observed the same behavioral effect that Kveraga et al. found, although in a slightly weaker form. One may speculate that Hick’s law applies for tasks in which a selection has to be made, while in our experiment responses appear to have been mostly reflexive in nature.

**Effects of background size and contiguity**

The apparent perceptual difference in stimulus presentation when going from two isolated squares (2-Image, Experiment 3, 17\(^{\circ}\) edge length, combined surface area of 578 deg\(^2\)) to a contiguous circle covering 19\(^{\circ}\) of visual field in all directions (Experiments 1 and 2, 38\(^{\circ}\) diameter, surface area of 1134 deg\(^2\)) is certainly not a trivial one; still, even though the active stimulus area increased by a factor of 2, there was no negative effect on rapid animal

<table>
<thead>
<tr>
<th>Target animals</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>(1st only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head visible</td>
<td>27.8%</td>
<td>10.3%</td>
<td>6.9%</td>
<td>5.8%</td>
<td>4.3%</td>
<td>45.4%</td>
</tr>
<tr>
<td>Head not visible</td>
<td>10.0%</td>
<td>9.6%</td>
<td>11.5%</td>
<td>6.8%</td>
<td>3.5%</td>
<td>25.8%</td>
</tr>
<tr>
<td>All</td>
<td>24.1%</td>
<td>10.2%</td>
<td>7.8%</td>
<td>6.0%</td>
<td>4.1%</td>
<td>41.4%</td>
</tr>
</tbody>
</table>

Table 10. Performance as predicted by low-level saliency.
detection performance or even a slight performance advantage. Given equal visibility, saccadic latencies have been shown to be shorter when target stimuli are presented on noisy backgrounds compared to uniform backgrounds (White, Stritzke, & Gegenfurtner, 2008). This result indicates that the circuits for driving saccades seem to be specifically tuned to vision in natural environments and may explain why the larger and more complicated natural scene background did not inhibit but even facilitate animal detection performance. Summarizing the absence of a performance impact for the large scene size together with the fact that the increase in target position uncertainty only had a tiny effect on saccade latency and no significant effect on (normalized) target accuracy, we conclude that our subjects most likely employed a parallel mode of visual search, in which the surface area and the number of candidate locations have little to no implications for the required processing time—at least up to the degree of uncertainty implemented in our experiments. Our subjects were not informed about the level of uncertainty in Experiment 1 (8-alternative), and the area covered by the targets sometimes overlapped between positions, making it harder to recognize any regularities in target placement; this suggests that visual search as performed by our subjects will extend to higher degrees of target location uncertainty, possibly encompassing the entire field of view. We cannot, however, fully exclude the possibility that our subjects accumulated some implicit knowledge about the number of available target locations during the experiment. The fact that Joubert et al. (2008) did find a decline in detection performance comparing 4 vs. 2 candidate locations may have been based on the fact that they used 4 distinct square-shaped images rather than one contiguous scene in their paradigm; we already found a significant decline in detection performance when changing from one contiguous scene to two individual squares, even though the underlying images and target position uncertainty were still the same. Registering 4 individual scenes, possibly very different from each other, may be a more challenging task to the visual system than the registration of one contiguous scene of comparable or even larger size. This would make sense if every scene was treated as a unit, rather than each unit area on the screen, and would be similar to the set size effects found in low-level psychophysical tasks (Eckstein, Thomas, Palmer, & Shimozaki, 2000; Palmer, 1994).

We see no reason to assume that a background context extending further in the periphery than the 19° covered by our stimuli would have any inhibiting effect—at least not in the type of experimental paradigm we employed. Thus, with the possible exception of the 3rd dimension, we believe that there is no obvious reason why this kind of rapid animal detection would be any less possible in full-field vision and, therefore, in real-life free viewing. This lends a high degree of validity to all previous studies utilizing either go/no-go or 2AFC paradigms.

**Target zones**

The design of our experiment allowed us to analyze how the subjects had detected the animal. Using the animal shapes as target areas, and recording the actual position of the saccade targets of our subjects, we were able to calculate the preferred saccade target zones on the animals. Our two candidate locations were the center of gravity (COG) of the animal and, alternatively, the animal’s head. Algorithmically, the center of gravity is a feature that can be computed in a simple and feed-forward fashion; it is also very robust to changes in image quality (e.g., resolution). At stimulus onset, our subjects are fixating at the screen center, and our target animal was always placed on a circle of 10.5° radius. This put the target well outside of the fovea, thereby limiting the acuity (resolution) available for detection and localization. The computation of the center of gravity as a principal feature of the animal would not have been affected by this decrease in resolution and would therefore have served as a reliable center of object to be used in the programming of the saccade. The computation of the center of gravity would, however, first require the entire animal to be segmented out from the image background, at least on a coarse scale. The head of the animal is usually much smaller than the shape of the animal, making it generally harder to detect; in many species, the head also shares features with the rest of the animal’s body—mostly color and texture (fur, feather, skin, etc.). This lets the detection of the head and the subsequent segmentation/separation of the head from the rest of the body appear more difficult than the computation of the center of gravity, especially in the periphery of the visual field. On the other hand, the head is probably the most telling feature of any animal. When encountering any animal in the real world, the head will give away information about the intentions of the animal (prey, flee, investigate, etc.), and it is generally the only part of an animal that will tell the observer whether he or she has been spotted by the animal. Additionally, the head may carry important information about the encountered species (e.g., the fangs of a predator). It would, therefore, certainly be useful for the observer to direct their eyes at the head of the animal as quickly as possible. From this point of view, it would be optimal if the observer’s gaze would saccade straight onto the head (with the first saccade). Due to the difficulty of correctly identifying the head of the animal in the visual periphery, a slower but more reliable strategy might be to saccade onto or near the animal with the first saccade, possibly using the center of gravity as an intermediate target, then performing a second saccade toward the head—which should be comparatively easy to locate once the observer’s gaze is already near the center of the animal. This would be consistent with the binary “zoom lens” strategy suggested by Zelinsky, Rao, Hayhoe, and Ballard (1997).
Effects of the presence of a visible head

While any visible animal always has a computable center of gravity, the head of the animal was hidden in some cases. Of the 294 images used in our experiments, 61 had no portion of the head visible—either due to occlusion by trees and other objects, or simply because the animal was turned the other way. The head of the animal was visible in 233 images, the overwhelming majority in views where at least some portion of the animal’s face was showing.

The presence of a visible head facilitated the rapid animal detection process, improving our subjects’ performance with respect to hit ratio (significant in Experiment 1 and Experiment 3; Experiment 2 was marginally not significant), hit latency (significant in all three experiments), and also decreasing the number of saccades needed to actually fixate on the animal shape. In the majority of all cases (68.3%), our subjects were able to reach the shape of the animal with the first saccade. This includes all images; for stimuli without a visible head, only 53.6% of the first saccades reached the animal. For stimuli with a visible head, 72.1% of the first saccades made it onto the animal shape. We thus conclude that the effect of a visible head on hit ratio, latency, and the number of saccades required to reach the animal shape is very pronounced. There is evidence that face detection in humans is even faster (as fast as 100–110 ms) than general animal detection (Crouzet, Kirchner, & Thorpe, 2010, in which underlying neuronal mechanisms are also discussed). In addition, face-selective neurons have been shown to respond in as little as 80–100 ms (Bruce, Desimone, & Gross, 1981; Jeffreys, 1996; Perrett, Rolls, & Caan, 1982). This indicates that it is actually possible to detect faces (or features thereof) in a time frame fast enough to be meaningful in the context of our experiments. Face detection might not be limited to human faces and therefore may also aid in the detection of animal faces, which in turn might facilitate animal detection. This could account for the difference in performance, both in speed and accuracy, between trials with and without visible heads.

We continued to analyze only the “with head” set of trials. In these, a trend emerged between the head and COG target zones: while the majority of all first saccades ending in either of the target zones was directed toward the COG (around 60%), this distribution evened out for the second saccades and was reversed for third saccades and later. Since second and third saccades generally originated from locations closer to the animal than the fixation point at trial start, the head position was probably easier to locate and might have been used as a target position more frequently because of this. As, however, in the vast majority of all trials the first saccade successfully entered the animal shape, the numerical magnitude of the second and, especially, the third (and later) saccades was comparatively small. Still, this trend was clearly recognizable in 8 of our 10 subjects, and one more subject was close to this trend (see Figure 11).

Low-level information cannot explain human performance

The hit ratio of simulated saccades based on low-level saliency was much lower than human performance; a task-independent model for bottom-up saliency can generally not be expected to match task-specific human performance (a more in-depth analysis of the performance of (saliency) map-based search algorithms can be found in Vincent, Troscianko, & Gilchrist, 2007). Still, saliency may explain a part of the general performance of our subjects, as the performance characteristics aside from the hit ratio showed: the increased performance for simulated saccades on animals with visible heads was similar to the tendency found in human observers, and hit ratio was somewhat reduced in the lower hemisphere, although the effect was not symmetric and did not quite match that found with humans. At this point, it is impossible to say to which extent human rapid animal detection is indeed based on low-level saliency; at the very least, there must be additional features that facilitate the detection process in human observers. This might be done in a multi-stage computation, in which candidate target locations that were possibly chosen based on low-level saliency are then further evaluated according to other criteria. Furthermore, the global amplitude spectra of the image classes used in this study contained no useful information as well, even though the presence of such information would have been expected to be of little use in human rapid animal detection (see Wichmann et al., 2010). Summarizing, our subjects were able to demonstrate a high level of performance, comparable to that of previous studies, without the help of dominant low-level cues.

Gaze strategy

Our data would be explained by the following strategy for the general detection and visual acquisition of animal targets.

Driven by the desire to acquire the animal’s head, a typical observer will saccade directly toward the head if the position of the head was detected with confidence. Otherwise, when the position of the head cannot be reliably determined, the observer will first approach the animal in a more robust way, possibly choosing the center of gravity as an alternative first target zone. Then, from close by, the observer will reevaluate the available visual information, hopefully being able to direct his gaze toward the head with greater confidence due to the increased acuity resulting from the reduced visual angle between the head and the fovea.
Acknowledgments

This work was supported by DFG Grant TR 528/1-4 (J. Drewes and J. Trommershaeuser). KG was supported by DFG Grant Ge 879/7. The collection of the natural images was supported by DFG Grants Wi 2103/1 and Ge 879/6 to Felix Wichmann and Karl Gegenfurtner. We would like to thank Natalie Wahl for her help in data collection.

Commercial relationships: none.
Corresponding author: Jan Drewes.
Email: mail@jandrewes.de.
Address: Centre de Recherche Cerveau et Cognition (UMR 5549), CNRS, Université Paul Sabatier, 133 Route de Narbonne, Fac de Medecine, Toulouse 31064, France.

References

All Natural Images Database (2011). Provided by the Institute of Experimental Psychology, Giessen University, Germany. Available from http://www.allpsych.uni-giessen.de/ANID.


