## The Time Course of Perceptual Grouping in Natural Scenes


Ilia Korjoukov, Danique Jeurissen, Niels A. Kloosterman, Josine E. Verhoeven, H. Steven Scholte and Pieter R. Roelfsema



*Psychological Science* 2012 23: 1482 originally published online 8 November 2012
DOI: 10.1177/0956797612443832

*Research Article*

Psychological Science
23(12) 1482–1489

DOI: 10.1177/0956797612443832
http://pss.sagepub.com

# The Time Course of Perceptual Grouping in Natural Scenes

**Ilia Korjoukov[1], Danique Jeurissen[1], Niels A. Kloosterman[2], Josine E. Verhoeven[1], H. Steven Scholte[2], and Pieter R. Roelfsema[1,3]**

[1]Department of Vision and Cognition, Netherlands Institute for Neuroscience, Royal Netherlands Academy of Arts and Sciences; [2]Department of Psychology, University of Amsterdam; and [3]Department of Integrative Neurophysiology, Centre for Neurogenomics and Cognitive Research, VU University Amsterdam

## Abstract

Visual perception starts with localized filters that subdivide the image into fragments that undergo separate analyses. The visual system has to reconstruct objects by grouping image fragments that belong to the same object. A widely held view is that perceptual grouping occurs in parallel across the visual scene and without attention. To test this idea, we measured the speed of grouping in pictures of animals and vehicles. In a classification task, these pictures were categorized efficiently. In an image-parsing task, participants reported whether two cues fell on the same or different objects, and we measured reaction times. Despite the participants' fast object classification, perceptual grouping required more time if the distance between cues was larger, and we observed an additional delay when the cues fell on different parts of a single object. Parsing was also slower for inverted than for upright objects. These results imply that perception starts with rapid object classification and that rapid classification is followed by a serial perceptual grouping phase, which is more efficient for objects in a familiar orientation than for objects in an unfamiliar orientation.

## Keywords

visual perception, attention, object recognition, perception

Received 9/15/11; Revision accepted 2/28/12

The human visual cortex starts its analysis of a visual scene with the extraction of low-level features, such as color, contour orientation, and spatial frequency; this process, which is performed by neurons with small receptive fields, occurs in parallel across the visual field. Psychologists call this feature-extraction phase *preattentive* because it is effortless and does not require attention (Julesz, 1981; Riesenhuber & Poggio, 1999a; Tovée, 1994). When people look around, they do not perceive a set of disconnected features, but rather experience coherent and unitary objects comprising many features; moreover, people are very apt in judging where in a picture one object ends and another one begins (Roelfsema & Houtkamp, 2011).

The mechanisms responsible for feature grouping and segregation are only partially understood. Some studies have found that perceptual grouping is a time-consuming (Jolicoeur, Ullman, & Mackay, 1986) and attention-demanding process (Ben-Av, Sagi, & Braun, 1992; Houtkamp, Spekreijse, & Roelfsema, 2003). For example, studies on a curve-tracing task—in which participants have to decide whether two cues are on the same curve or different curves—have shown that reaction times (RTs) increase linearly with the distance between these cues and that grouping is both time-consuming and attention demanding (Jolicoeur et al., 1986; Pringle & Egeth, 1988). In this task, perceptual grouping of contour elements is associated with the gradual spread of object-based attention over the curve (Houtkamp et al., 2003).

These findings contrast with the popular view that perceptual grouping is highly efficient and does not require attention (Julesz, 1981; Treisman & Gelade, 1980). One mechanism that could produce efficient perceptual grouping is the convergence of visual attributes onto single neurons in higher areas in the visual cortex (Riesenhuber & Poggio, 1999b; Roelfsema, 2006; Tovée, 1994). These neurons are selective for object shape, which implies that they are tuned to groups of low-level features in specific spatial configurations (Hung, Kreiman, Poggio, & DiCarlo, 2005; Tanaka, 1993). Thorpe and his colleagues (Kirchner & Thorpe, 2006; Thorpe, Fize, & Marlot, 1996) demonstrated that observers are indeed very efficient in

**Corresponding Author:**
Pieter R. Roelfsema, Department of Vision and Cognition, Netherlands Institute for Neuroscience, Amsterdam, The Netherlands
E-mail: p.roelfsema@nin.knaw.nl

recognizing members of object categories, such as animals or vehicles, even if the objects appear in complex scenes. Moreover, in many (Li, VanRullen, Koch, & Perona, 2002; Peelen, Fei-Fei, & Kastner, 2009) but not all (Walker, Stafford, & Davis, 2008) situations, attention appears to be unnecessary for detection of these object categories. If object categorization is so efficient, why have other studies consistently found delays during the grouping of low-level image elements? Are the delays observed in curve tracing a curiosity of the artificial task, or do they also occur in natural viewing conditions?

We investigated the possibility that these apparent discrepancies among studies are caused by differences between the mechanisms for object recognition and image parsing (Peterson, Harvey, & Weidenbacher, 1991; Roelfsema, 2006; Vecera & Farah, 1997). The detection of object categories may be realized by fast feed-forward processing in shape-selective areas of the visual cortex, whereas the assignment of features and image regions to a specific object may require additional processing. For example, if there are two animals in a picture, the many animal features, such as eyes and paws, make animal detection easy and efficient, but additional processing may be required to group the features of one animal together and to segregate them from the features of the other one. For the study reported here, we devised a new task to measure the processing delays in an explicit image-parsing task, and we compared them with the delays that occur in picture categorization. If shape recognition precedes image parsing, then manipulations that impair recognition might also delay parsing. We tested this prediction by varying picture orientation, because inverted pictures might be associated with a protracted parsing process. The results show that image parsing in natural images is a serial process that benefits from the outcome of a preceding object recognition stage.

## Experiment 1

This experiment tested the time course of explicit perceptual grouping in natural images. We presented pictures with two animals or vehicles and asked observers to indicate whether two cues fell on the same object or different objects (Fig. 1a). If perceptual grouping of natural objects is a parallel process, then RT should be independent of the distance between the image elements that have to be grouped. However, if perceptual grouping invokes a serial process, then participants' RT should increase with the distance between image elements, as in the case of contour-group tasks, such as curve tracing (Egeth & Yantis, 1997; Jolicoeur et al., 1986).

### Method

***Participants.*** Twenty people (5 males, 15 females; 3 left-handed, 17 right-handed) participated in this experiment. Their mean age was 19.9 years (range = 18–24). The participants had normal or corrected-to-normal visual acuity and were paid for their participation. Ethical approval was obtained through the Psychonomic Ethics Committee at the University of Amsterdam.

***Stimuli.*** Each trial began with the presentation of a white fixation point (size: 0.2° visual angle) for 500 ms on a black screen (Fig. 1a). Next, two white cues were shown for 1,000 ms. Finally, a picture with two animals or vehicles appeared with the two cues (cue size: 0.2°); the cues were superimposed and flickered at a frequency of 10 Hz to guarantee their visibility. The pictures had a size of 34° × 25° and were centered on the black background, such that there was a black border around each picture. The picture and cues stayed on the screen until the participants gave a response or 5,000 ms had elapsed. Visual feedback ("correct" or "incorrect") was provided at the end of every trial.

For a given picture, one cue appeared in a position that was fixed for that particular picture (cue *f* ), whereas the second cue was in one of six other positions (which defined six cue-location conditions; see Fig. 1b). Three locations were on the same object as *f* (*s* trials; 50%), and three were on a different object from *f* (*d* trials; the other 50%). On *s* trials, the second cue was on the same part of the object as *f* (e.g., both cues on the body of a tiger) but separated from *f* by a distance of 6° or 12°, or was on a different part of the object (e.g., on the head of the tiger) and separated from *f* by 12° ($s_1$, $s_2$, and $s_3$ cues, respectively, in Fig. 1b). These same three distances were used on *d* trials (separation of 6° for $d_1$ and 12° for $d_2$ and $d_3$). Twenty-four pictures were used for the experiment (all pictures are shown in Fig. S1 in the Supplemental Material available online). The pictures were created from high-resolution color images obtained from online open sources. Twelve pictures contained two animals, and the other 12 contained two vehicles. The size of each animal or vehicle was comparable to that of the other in the pair, and they occupied a significant fraction of the foreground. We used a standard graphical editor that autocorrected each picture's luminance contrast and white balance

***Procedure.*** During the experiment, participants sat in a dimly lit room with the head supported by a chin rest at a distance of 53 cm from a 19-in. CRT monitor (1024 × 768 pixels; 100-Hz frame rate) controlled through a PC (Windows-controlled Dell computer). After onset of each picture, participants indicated whether the two cues were located on the same object or on different objects, by pressing the "z" or "/" key on a keyboard. The assignment of keys was counterbalanced across participants. The participants first performed 8 training trials with a separate set of similar stimuli. Accuracy was emphasized in the instructions, but the participants were also asked to respond quickly. Participants completed six blocks of 144 trials, with short breaks between blocks. Within a block, the six cue-location conditions for each of the 24 stimuli were presented in a random order. Six trials per condition and per picture were presented so that RTs could be assessed for each individual picture.

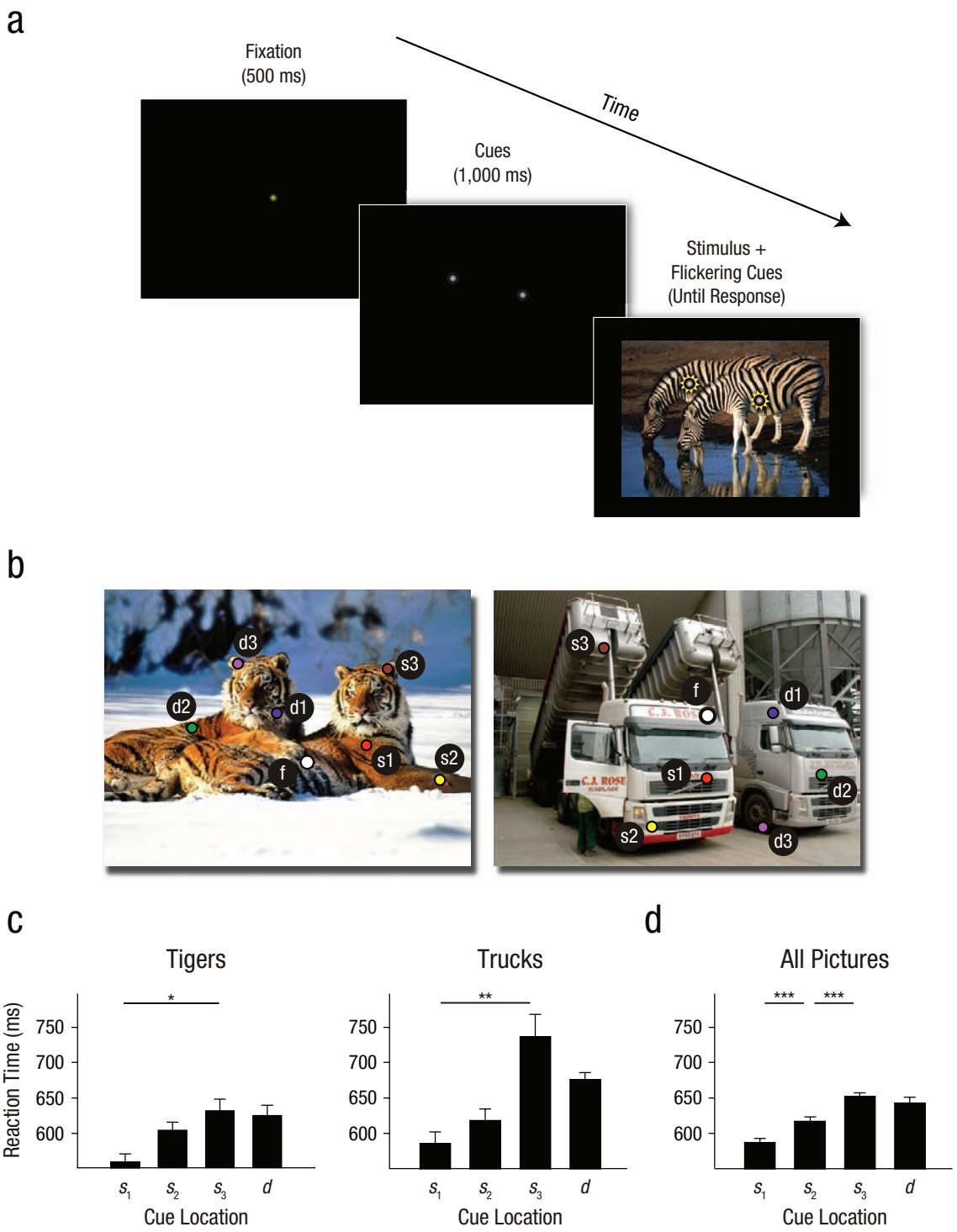


**Fig. 1.** Paradigm and results for Experiment 1. At the beginning of each trial (a), participants saw a fixation point, which was followed by two cues. After an additional delay, a picture with two animals or two vehicles appeared while the cues remained on-screen. Participants were asked to indicate whether the flickering cues fell on the same object or on different objects. In (b), the dots (colored here for visualization purposes) denote the possible cue locations. For a given picture, one of the cues was always in the same location (cue *f* ), whereas the other cue appeared in one of six possible locations, either on the same object as cue *f* (cues labeled *s*) or on the other object (cues labeled *d*). The graphs in (c) show mean reaction time (RT) as a function of cue location for the picture with the tigers and the picture with the trucks. The graph in (d) shows mean RT averaged across all participants and all pictures. In all graphs, asterisks denote significant differences between cue locations, as determined with post hoc tests (*$p < .05$, **$p < .01$, ***$p < .001$); note that RTs for all *d* cue locations were averaged together. Error bars denote standard errors of the mean after subtraction of the participants' overall mean RTs.

***Data analysis.*** All trials with RTs shorter than 300 ms or longer than 3,000 ms were removed from the data set (< 1% of the trials). A two-way repeated measures analysis of variance (ANOVA) with picture number and cue location ($s_1$, $s_2$, $s_3$, or $d$ ) as factors was used to test differences in RTs. Only RTs for correct responses were analyzed. Differences between conditions were further analyzed with planned pairwise comparisons. Greenhouse-Geisser correction was performed when necessary. For our main statistical analysis, we used arithmetic means, but we obtained similar results when we analyzed harmonic means or medians (see Supplementary Information, including Table S1, in the Supplemental Material). Generalization to other picture sets was tested with the minimum value of the quasi $F$ test (Clark, 1973).

## Results and discussion

The participants achieved a mean accuracy of 92.4%. The accuracies for cues $s_1$, $s_2$, and $s_3$ (on the same object) were 92.6%, 92.0%, and 91.5%, respectively, and accuracy was 93.3% for $d$ trials: A one-way repeated measures ANOVA, with $s_1$, $s_2$, $s_3$, and $d$ as factors, indicated that the differences in accuracy across cue locations were not significant, $F(3, 57) = 1.9$, $p > .15$. Figure 1c illustrates how RT depended on the position of the cues for the picture with two tigers. RT was shortest for cue $s_1$, which was nearest to $f$; it increased by 45 ms for $s_2$ and by 73 ms for $s_3$ (compared with $s_1$; recall that $s_3$ was on a different part, the head of the tiger). Thus, if the two cues fell on the same animal, RTs increased with distance, and they were particularly long if the features belonged to different parts of the animal. Similar results were obtained for a picture with two trucks (Fig. 1c) and also for many other pictures (see Fig. S1). Mean RT (Fig. 1d) was 590 ms for the shortest distance on the same object ($s_1$) and increased by 29 ms for the next larger distance on the same object ($s_2$) and by 66 ms (compared with $s_1$) when the second cue was on a different part of the object ($s_3$); the main effect of cue location for $s$ cues was significant, $F(1.9, 35.6) = 22.7$, $p < .00001$, Greenhouse-Geisser corrected. The RT on $d$ trials was on average 23 ms longer than the RT on $s$ trials, $F(1, 19) = 4.91$, $p < .05$ (planned comparison).

Would the effect of cue location on RT also occur in a different set of pictures with similar properties? To assess the generality of the effect of cue position on RT, we estimated the minimum value of the quasi $F$ statistic (Clark, 1973), which was significant, minimum $F'(3, 91) = 5.46$, $p < .01$. Thus, the influence of cue position on RT remained significant if both participants and pictures were treated as random factors. There was no speed-accuracy trade-off, because conditions in which longer RTs were observed on $s$ trials were not associated with lower error rates.

We next investigated whether our use of a fixed set of images that were seen repeatedly induced learning, using a repeated measures ANOVA with block and cue location as factors. We observed a general speeding of RTs during the experiment, as the mean RT was 682 ms in the first block of trials and decreased to 608 ms in the last block. However, experience with the task did not influence the differences in RT between cue-location conditions as there was no interaction between condition and block number, $F(25, 475) = 0.97$, $p = .51$. A previous study (Kirchner & Thorpe, 2006) showed that learning did not reduce RTs in an image categorization task in which participants saw two pictures, one on the left and one on the right, and were asked to make saccadic eye movements toward the picture containing an animal. We therefore suggest that the shortening of RT over time in our experiment may have been caused by participants' learning to map "same" and "different" responses onto the responses buttons.

Our results suggest that the parsing of natural images calls on a serial process with a processing time that increases if distance increases, and if the to-be-grouped image elements belong to different parts of the object. We considered the possibility that our results were influenced by eye movements. Eye movements between cues that were farther apart may have caused longer delays in parsing. Therefore, using an eye tracker to monitor eye position, we repeated the experiment while participants maintained fixation on a fixation point. We obtained virtually identical results in this control experiment (see Supplementary Information and Fig. S2 in the Supplemental Material), which ruled out eye movements as a cause of the processing delays. Another control experiment (see Supplementary Information and Fig. S2) ruled out picture size as a cause of the processing delays.

We conclude that the grouping of parts that belong to the same object is associated with serial processing. RTs in the task were on the order of 600 ms, which is relatively long. In Experiment 2, we tested whether the serial process responsible for the grouping of image elements occurs after object categorization.

# Experiment 2

The second experiment compared RTs in the image-parsing task with RTs in an object classification task. If image parsing depends on object categorization, then parsing might be more efficient for pictures in a familiar configuration. To influence the efficiency of object categorization and image parsing, we varied picture orientation (upright vs. inverted). If image categorization precedes parsing, categorization results become available for parsing, and shorter processing times would be expected for upright pictures, whose parts are in the expected configuration; parsing should be less efficient for upside-down pictures (Peterson et al., 1991; Vecera & Farah, 1997). The influence of low-level grouping cues, such as good continuation, similarity, and connectedness, is not expected to depend on picture orientation.

## Method

***Participants.*** Twenty-four people with normal or corrected-to-normal vision (4 males, 20 females; 1 left-handed, 23 right-handed) participated in this experiment. Their mean age was

20.9 years (range = 18–25), and they were paid €10 per hour for their participation.

***Stimuli and procedure.*** Participants were assigned to either the categorization task (12 participants) or the image-parsing task (12 participants). They performed eight training trials. The pictures were the same as in Experiment 1. Each trial began with the presentation of a white fixation point for 500 ms on a black screen. In the categorization task, a picture appeared next and remained on-screen until the response. In the image-parsing task, after the fixation display, the cues were shown for 1,000 ms before the picture appeared with the superimposed cues. Participants in the categorization task indicated whether the picture contained animals or vehicles by pressing the "z" or "/" key on a keyboard; the assignment of keys was counterbalanced across participants. Participants in the image-parsing task used these keys to indicate whether the two cues were on the same or different animals or vehicles. In both tasks, visual feedback was given at the end of every trial. Blocks with pictures in an upright orientation (50% of blocks) were interleaved with blocks with inverted pictures (the other 50%), and the order of these blocks was counterbalanced across participants. Participants were instructed to respond quickly, but accuracy was emphasized. As in Experiment 1, participants completed six blocks of 144 trials each.

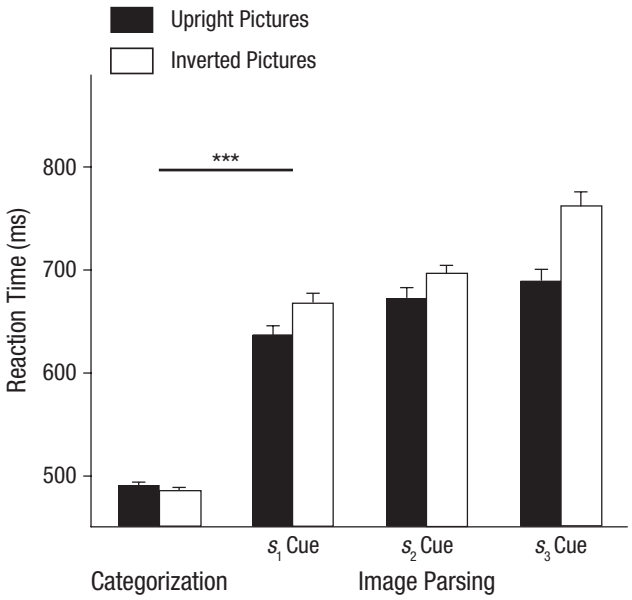


**Fig. 2.** Results for Experiment 2: mean reaction time (RT) as a function of task and picture orientation. For the image-parsing task, results are shown separately for the three different positions of the *s* cue (i.e., cues located on the same object as *f*). Error bars denote standard errors of the mean after subtraction of the participants' mean RT. Asterisks denote a significant difference between RT in the classification task and RT in the fastest cue condition ($s_1$) of the image-parsing task (***$p < .001$).

## Results

In the image-categorization task, accuracy (96% for upright pictures vs. 97% for inverted pictures) and RT (490 vs. 485 ms; see Fig. 2) were similar for upright and inverted pictures, $p > .5$ and $p > .3$, respectively; this finding is in accordance with that of a previous study (Rousselet, Macé, & Fabre-Thorpe, 2003). These short RTs show that participants were quickly able to decide whether a picture contained an item of a specific category.

In the image-parsing task, accuracy for inverted pictures and accuracy for upright pictures were both 91%. The mean accuracies for cue locations $s_1$, $s_2$, and $s_3$ were 92.5%, 90.3%, and 90.3%, respectively; differences in accuracy were not significant, $F(2, 22) = 2.4$, $p > .1$. Figure 2 shows the RTs in the image-parsing task. The mean RT for cue $s_1$ (the fastest condition in the image-parsing task) was 656 ms, which was 168 ms longer than the mean RT in the classification task, $F(1, 22) = 20.43$, $p < .001$. This implies that participants could classify the pictures well before they responded in the image-parsing task. Experiment 2 reproduced the pattern of RTs of Experiment 1, as RTs increased if the cues were farther apart on the same object and were even longer if the cues fell on different parts of the same object; the main effect of cue location was significant, $F(2, 22) = 22.9$, $p < .00001$. This effect was also significant when we treated the picture as a random factor (Clark, 1973), minimum $F'(2, 52) = 9.8$, $p < .001$. The increase in RT was not caused by a speed-accuracy trade-off, because accuracy did not improve significantly in the slowest conditions, $s_2$ and $s_3$.

Although categorization speed did not depend on picture orientation, image parsing took more time for inverted pictures than for upright pictures, $F(1, 11) = 13.6$, $p < .01$; minimum $F'(1, 25) = 8.08$, $p < .01$. We observed an increase in RT for inverted pictures compared with upright pictures in all cue locations: 36 ms and 24 ms for locations $s_1$ and $s_2$, respectively, and 83 ms for $s_3$, when the cue was located on a different part of the object (Fig. 2). The interaction between cue location and picture orientation was significant, $F(2, 22) = 7$, $p < .01$. These results imply that image parsing continues after object categorization in natural scenes and that categorization results aid in the parsing process.

## General Discussion

A picture is initially represented in a highly fragmented manner by neurons distributed across many areas of the visual cortex. We investigated the time course of the perceptual organization processes that impose structure on such distributed representations for natural images. We found that perceptual grouping invokes a serial process that takes longer for elements farther apart and even more time for elements on different parts of an object. A control experiment demonstrated that these delays were not due to eye movements.

At first thought, it may seem counterintuitive that parsing is a serial process given the popular view that object recognition is a fast and efficient process. However, when we compared processing times in the image-parsing task with those in the picture categorization task, we found that parsing continues after categorization. This result is in accordance with previous work suggesting that object recognition can provide useful information for image parsing (Peterson et al., 1991; Vecera & Farah, 1997). We found that the speed of object categorization did not depend on picture orientation (Rousselet et al., 2003), but that image parsing was slower for inverted pictures than for upright pictures, and that parsing speed decreased further for elements on different parts of the inverted object. Thus, perceptual grouping makes use of the outcome of a successful object recognition process. This finding supports models showing that feedback from shape-selective representations can guide the grouping of low-level features at earlier processing levels (Sharon, Galun, Sharon, Basri, & Brandt, 2006; Tsotsos, Rodríguez-Sánchez, Rothenstein, & Simine, 2008; Vecera & O'Reilly, 1998).

These results support the view that there is a fast, feed-forward process, as well as serial, recurrent processes, for perceptual grouping (Roelfsema, 2006). The feed-forward process can explain why object recognition and categorization are highly efficient under some conditions. Thorpe et al. (1996) showed that the detection of specific object categories, such as animals and vehicles, can be completed within less than 150 ms. This fast object categorization process can rely on features of intermediate complexity (Ullman, Vidal-Naquet, & Sali, 2002), such as the shape of eyes and paws for animals and of tire-covered wheels and steering wheels for vehicles. Neurons in the inferotemporal cortex of monkeys respond selectively to these feature constellations (Tanaka, 1993), and this selective response implies that some feature combinations can be quickly detected by the convergence of visual attributes onto single neurons (Riesenhuber & Poggio, 1999b; Roelfsema, 2006). Information about these feature constellations is present in the earliest part of the neuronal responses, in accordance with a fast, feed-forward processing phase (Hung et al., 2005) that corresponds to preattentive vision (Julesz, 1981; Peelen et al., 2009; Riesenhuber & Poggio, 1999a; Tovée, 1994; Treisman & Gelade, 1980).

Some tasks are solved as soon as objects have been detected or categorized. However, there are many other tasks that depend on successful image parsing and have to rely on the later, serial processing phase. For example, it is crucial to detect and group all parts of an object if one wants to grasp it, and the same holds true if one wants to avoid collisions of the grasped object with other objects. The present results do not imply that categorization always precedes parsing. Parsing may precede recognition, for example, if the object is camouflaged, and this variation in the processing order implies reciprocal interactions between parsing and recognition (Roelfsema & Singer, 1998; Vecera & O'Reilly, 1998). In accordance with this view, interference created by transcranial magnetic stimulation of early visual areas during a phase when the higher visual areas have become active can still impair picture categorization (Koivisto, Railo, Revonsuo, Vanni, & Salminen-Vaparanta, 2011).

The present results demonstrate that the explicit grouping of image elements in natural images requires serial processing. This serial grouping process is reminiscent of contour-grouping (curve-tracing) tasks in which participants indicate whether two cues fall on the same curve or on different curves (Fig. 3a). RT in these tasks increases linearly with the distance between the two cues as measured along the same curve (Jolicoeur et al., 1986; Jolicoeur, Ullman, & MacKay, 1991), and participants gradually spread object-based attention from one contour element to the next until the curve is entirely labeled by attention (Fig. 3b; see also Houtkamp et al., 2003). This labeling process is implemented in the visual cortex as the gradual propagation of enhanced neuronal activity along the relevant curve (Roelfsema, 2006; Roelfsema, Lamme, & Spekreijse, 1998, 2004), a process that can also be measured by electroencephalography (Lefebvre, Jolicoeur, & Dell'Acqua, 2010). Comparable delays also occur for other Gestalt grouping cues—such as proximity, similarity, and common fate—that promote perceptual grouping between adjacent image elements (Houtkamp & Roelfsema, 2010) and determine the spread of attentional response modulation in the visual cortex (Wannig, Stanişor, & Roelfsema, 2011). In the present experiment, we observed equivalent delays in the grouping of image elements of objects in natural scenes (Figs. 3c and 3d). The typical delays ranged from 30 to 60 ms, and this range is similar to that of delays observed with short curves in the curve-tracing task (Pringle & Egeth, 1988), although we did observe longer delays for specific pictures (see Fig. S1).

The similarity between natural image parsing and curve tracing suggests that the parts of objects in natural scenes are also grouped once those parts are labeled with object-based attention (Roelfsema & Houtkamp, 2011; Figs. 3c and 3d). We presume that attention spreads according to Gestalt grouping cues, a process that may be implemented in early visual cortex (Ben-Shahar, Scholl, & Zucker, 2007; Bhatt, Carpenter, & Grossberg, 2007; Roelfsema, 2006; Wannig et al., 2011). This incremental grouping process is complete once all image elements of an object are indexed by object-based attention as a "grouped array" (Vecera & Farah, 1994; see also Driver, Davis, Russell, Turatto, & Freeman, 2001; Duncan, 1984). The extra delays that occur during parsing of inverted pictures suggest that object recognition augments this attention-spreading process by providing information about the typical configuration of object parts.

In summary, our results support the idea that perceptual grouping starts with the preattentive extraction of low-level features and features of intermediate complexity and culminates in the detection of object categories. This feed-forward processing phase is followed by a serial image-parsing phase in which object-based attention indexes the set of low-level and high-level features that belong to a unitary perceptual object. An exciting idea inspired by these findings is that the serial operations used for contour grouping (see Fig. 3) are also important for the perception of everyday scenes.

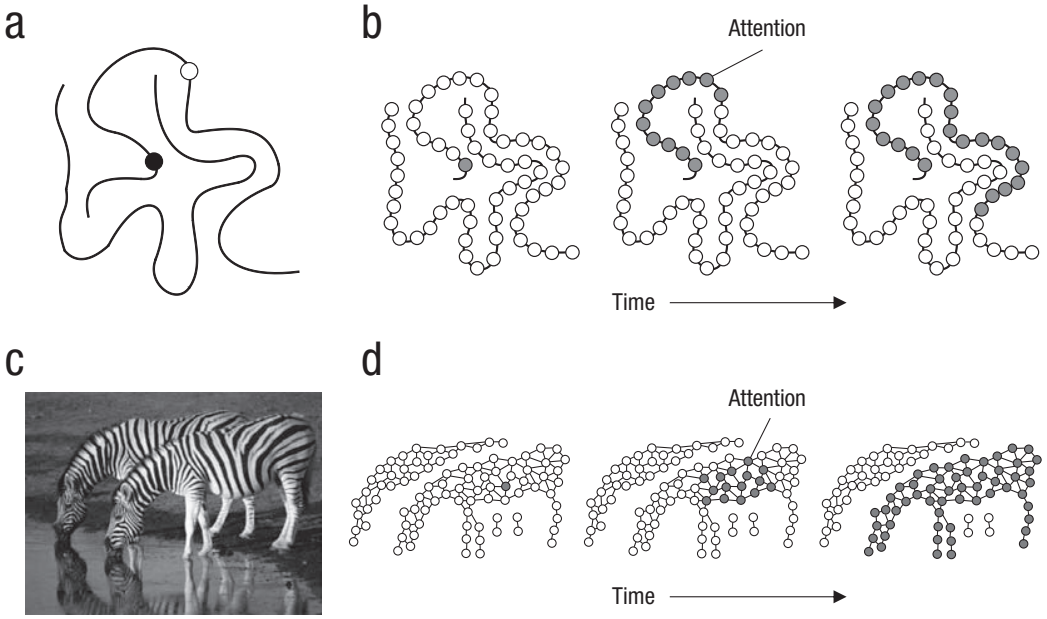


**Fig. 3.** Illustration of the serial grouping process in the curve-tracing and natural-image-parsing tasks. In the curve-tracing task, participants indicate whether two cues (black and white dots) fall on the same curve or on different curves (a). The task is accomplished by spreading object-based attention (gray circles) from one contour element to the next until a given curve is entirely labeled by attention (b). The image-parsing task uses natural scenes (c). The task is similar to the curve-tracing task, as participants indicate whether two cues fall on the same object. Object-based attention spreads across the surface of a given object in order to group the image as one object (d).

## Acknowledgments

The authors thank Mary Peterson and Jan Theeuwes for useful comments on an earlier version of the manuscript.

## Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

## Funding

This work was supported by grants to P. R. R. from three programs of The Netherlands Organization for Scientific Research (NWO): NWO-Exact, NWO Maatschappij- en Gedragswetenschappen (MaGW), and NWO Vici.

## Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

## References

Ben-Av, M. B., Sagi, D., & Braun, J. (1992). Visual attention and perceptual grouping. *Perception & Psychophysics*, *52*, 277–294.

Ben-Shahar, O., Scholl, B. J., & Zucker, S. W. (2007). Attention, segregation, and textons: Bridging the gap between object-based attention and texton-based segregation. *Vision Research*, *47*, 845–860.

Bhatt, R., Carpenter, G. A., & Grossberg, S. (2007). Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research*, *47*, 3173–3211.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–358.

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention and phenomenal visual objects. *Cognition*, *80*, 61–95.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*, 501–517.

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology*, *48*, 269–297.

Houtkamp, R., & Roelfsema, P. R. (2010). Parallel and serial grouping of image elements in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1443–1459.

Houtkamp, R., Spekreijse, H., & Roelfsema, P. R. (2003). A gradual spread of attention during mental curve tracing. *Perception & Psychophysics*, *65*, 1136–1144.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*, 863–866.

Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition*, *14*, 129–140.

Jolicoeur, P., Ullman, S., & MacKay, M. (1991). Visual curve tracing properties. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 997–1022.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature*, *290*, 91–97.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*, 1762–1776.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural sciences. *Journal of Neuroscience*, *31*, 2488–2492.

Lefebvre, C., Jolicoeur, P., & Dell'Acqua, R. (2010). Electrophysiological evidence of enhanced cortical activity in the human brain during visual curve tracing. *Vision Research*, *50*, 1321–1327.

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

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*, 94–98.

Peterson, M. A., Harvey, E. M., & Weidenbacher, H. J. (1991). Shape recognition contributions to figure-ground reversal: Which route counts? *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 1075–1089.

Pringle, R., & Egeth, H. E. (1988). Mental curve tracing with elementary stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 716–728.

Riesenhuber, M., & Poggio, T. (1999a). Are cortical models really bound by the "binding problem"? *Neuron*, *24*, 87–93.

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

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, *29*, 203–227.

Roelfsema, P. R., & Houtkamp, R. (2011). Incremental grouping of image elements in vision. *Attention, Perception, & Psychophysics*, *73*, 2542–2572.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*, 376–381.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nature Neuroscience*, *7*, 982–991.

Roelfsema, P. R., & Singer, W. (1998). Detecting connectedness. *Cerebral Cortex*, *8*, 385–396.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision*, *3*(6), Article 5. Retrieved from http://journalofvision.org/content/3/6/5

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

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, *262*, 685–688.

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

Tovée, M. J. (1994). How fast is the speed of thought? *Current Biology*, *4*, 1125–1127.

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

Tsotsos, J. K., Rodríguez-Sánchez, A. J., Rothenstein, A. L., & Simine, E. (2008). The different stages of visual recognition need different attentional binding strategies. *Brain Research*, *1225*, 119–132.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*, 682–687.

Vecera, S. P., & Farah, M. J. (1994). Does visual attention select objects or locations? *Journal of Experimental Psychology: General*, *123*, 146–160.

Vecera, S. P., & Farah, M. J. (1997). Is visual image segmentation a bottom-up or an interactive process? *Perception & Psychophysics*, *59*, 1280–1296.

Vecera, S. P., & O'Reilly, R. C. (1998). Figure-ground organization and object recognition processes: An interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 441–462.

Walker, S., Stafford, P., & Davis, G. (2008). Ultra-rapid categorization requires visual attention: Scenes with multiple foreground objects. *Journal of Vision*, *8*(4), Article 21. Retrieved from http://journalofvision.org/content/8/4/21

Wannig, A., Stanişor, L., & Roelfsema, P. R. (2011). Automatic spread of attentional response modulation along gestalt criteria in primary visual cortex. *Nature Neuroscience*, *14*, 1243–1244.

**Supplementary information for "The time-course of perceptual grouping in natural scenes"**

*Ilia Korjoukov, Danique Jeurissen, Niels A. Kloosterman, Josine E. Verhoeven, H. Steven Scholte &, Pieter R. Roelfsema*

We performed two additional control experiments to test how picture size and eye-movements influence the reaction time in the image-parsing experiment. Participants carried out the image-parsing task of Experiment 1 (main text) with some changes. One group of saw smaller pictures (Size experiment) and another group of participants saw the original pictures but they had to maintain their gaze on a central fixation point while eye position was monitored with an eye-tracker (Fixation experiment).

***Methods***

*Participants*

A total of thirty participants (17 males, 4 right-handed) participated in these two experiments. Their average age was 26 years, they had normal or corrected-to-normal acuity and were paid for their participation. Of the 30 participants, 18 participated in the Size experiment and 12 participated in the Fixation experiment. Ethical approval was obtained through the Psychonomic Ethics Committee at the University of Amsterdam.

*Stimuli and Procedure*

Participants started a trial by fixating on the white central fixation point for 700 ms on a black screen. Then the second white cue appeared for 1000 ms. The picture with two animals or vehicles was shown with the cues superimposed and flickering at a frequency of 10 Hz. The picture and cues were shown until the participant gave a response or 5000 ms had elapsed. Visual feedback (correct, error, too late or fixation break) was provided at the end of every trial. The instructions emphasized accuracy and participants completed six blocks of 144 trials each.

In the Size experiment, the viewing distance was increased to 90 cm, no chinrest was used, and the responses were given by pressing one of two buttons located on both sides of a chair. The experiment was run with simultaneous EEG recording (the EEG data is not presented here). Due to the increase in viewing distance, the picture size was reduced 1.8 times (to 19x14 degrees) compared to Experiment 1.

In the Fixation experiment, the task was the same as in the Experiment 1, however, the stimulus layout on the screen was different. Point *f* (Figure 1 and Figure S1) coincided with the fixation point and was always presented in the center of the screen (position of the pictures on the screen therefore differed). We scaled the pictures to 30 x 22 degrees (corresponding to 87% of the size in Experiment 1) so that all shifted pictures fitted on the screen. The relative positions of the other cues ($s_1$, $s_2$, $s_3$, *d*) and their distances to *f* were preserved. The participant was required to maintain fixation throughout the trial. Whenever the eye-position deviated by more than 1.5 degrees from the fixation point, the trial was aborted and repeated later in the same block. Gaze position was recorded at a sampling rate of 250Hz with an EyeLink eye-tracking system (SR Research Ltd).

**Results**

Trials with RTs shorter than 300 ms or longer than 3000 ms were removed from the data (less than 1% of the trials). The overall accuracy was 95% in the Size experiment and 89% in the Fixation experiment. The accuracy tended to decrease for conditions with a longer RT (no evidence for a speed-accuracy tradeoff).

*Size experiment:* Although the overall reaction times were longer than in the main experiment the influence of cue position was similar. Average RTs were 677 ms for correct trials with the shortest distance $s_1$, they increased by 59 ms for $s_2$, and by 95 ms for $s_3$, on a different part of the object. An ANOVA with cue-distance as factor revealed a significant main effect of the cue-distance ($F_{3,51}$=26.4, $P<10^{-5}$, Greenhouse-Geisser corrected, min $F'_{3,118}$=9.5,$P<10^{-4}$).

*Fixation experiment:* Again, the influence of the cue position on RT was similar to that in Experiment 1. The RTs for $s_1$ were 660 ms, on average, they increased by 39 ms for $s_2$, and by 93 ms for $s_3$. An ANOVA revealed a significant main effect of the cue-distance ($F_{3,33}$=15.2, $P<10^{-5}$, Greenhouse-Geisser corrected, min $F'_{3,84}$=8.2,$P<10^{-4}$).

*Comparison across experiment*s: We next carried out a repeated-measures two-way ANOVA to compare the cue-position effect across experiments, with one between-subject factor (experiment number) and one within subject factor (cue: $s_1$, $s_2$, $s_3$, $d$). As expected, we found a significant main effect of the cue-distance ($F_{3,141}$=63.2, $P$<$10^{-6}$) that did not interact with experiment ($F_{6,141}$=1.7, $P$>0.1) indicating the effect of cue position on RT was similar across experiments (see Fig. S2).

**Discussion**

The control experiments replicated the findings from Experiment 1 for smaller stimuli and with restricted eye-movements. Because the speed of image-parsing in the two control experiments is similar to that in Experiment 1, we conclude (1) that eye-movements cannot explain the observed pattern of processing delays in Experiment 1 and (2) that the results are robust across substantial changes in the size of the pictures. The observed processing delays appear to be scale invariant; they stay constant when parsing occurs over smaller distances in smaller pictures. This scale invariance is reminiscent of previous results with a curve tracing task where overall changes in the size of stimuli had little influence on the RTs (Jolicoeur and Ingleton, 1991).

**References**

Jolicoeur,P., and Ingleton,M. (1991). Size invariance in curve tracing. Mem. Cognit. *19*, 21-36.

| Experiment | Averaging method | s1 ±SE | s2 ±SE | s3 ±SE | d ±SE | F | $df_t$ / $df_e$ | P |
|---|---|---|---|---|---|---|---|---|
| Experiment 1 (image parsing) | Arithmetic Mean | 590 ±5 | 619 ±4 | 656 ±4 | 644 ±7 | 22.7 | 1.9/35.6 | $<10^{-5}$ |
| | Harmonic Mean | 562 ±4 | 589 ±3 | 619 ±4 | 601 ±7 | 20.7 | 1.7/33.2 | $<10^{-5}$ |
| | Median | 564 ±5 | 586 ±3 | 620 ±4 | 596 ±6 | 18 | 2/39 | $<10^{-5}$ |
| Experiment 2 (inverted pictures) | Arithmetic Mean | 656 ±7 | 686 ±7 | 735 ±10 | 707 ±12 | 22.0 | 2/22 | $<10^{-5}$ |
| | Harmonic Mean | 617 ±5 | 644 ±4 | 687 ±8 | 654 ±11 | 35.8 | 2/22 | $<10^{-6}$ |
| | Median | 621 ±6 | 648 ±4 | 700 ±9 | 656 ±10 | 32.5 | 2/22 | $<10^{-6}$ |
| Suppl. Exp 1 (size control) | Arithmetic Mean | 677 ±9 | 736 ±4 | 772 ±7 | 742 ±6 | 26.4 | 2.0/33.4 | $<10^{-6}$ |
| | Harmonic Mean | 640 ±8 | 693 ±3 | 722 ±6 | 692 ±6 | 27 | 2.1/35.1 | $<10^{-7}$ |
| | Median | 639 ±9 | 694 ±3 | 725 ±8 | 692 ±7 | 19 | 1.9/33 | $<10^{-5}$ |
| Suppl. Exp 2 (fixation control) | Arithmetic Mean | 660 ±8 | 699 ±4 | 753 ±9 | 705 ±11 | 15.2 | 2/22.1 | $<10^{-4}$ |
| | Harmonic Mean | 638 ±7 | 675 ±4 | 718 ±7 | 678 ±11 | 13.2 | 1.8/20 | $<10^{-3}$ |
| | Median | 638 ±7 | 681 ±5 | 724 ±9 | 673 ±13 | 12.3 | 1.7/18.6 | $<10^{-3}$ |

**(Table S1 continued)**

| Experiment | Averaging method | $MS_t$ / $MS_e$ | Min F' | $df'_t$ / $df'_e$ | P' |
|---|---|---|---|---|---|
| Experiment 1 (image parsing) | Arithmetic Mean | 16913/746 | 5.5 | 3/91 | $<10^{-2}$ |
| | Harmonic Mean | 11443/553 | 8.6 | 3/125 | $<10^{-4}$ |
| | Median | 10795/600 | 6.8 | 3/123 | $<10^{-3}$ |
| Experiment 2 (inverted pictures) | Arithmetic Mean | 19288/841 | 9.8 | 2/52 | $<10^{-3}$ |
| | Harmonic Mean | 14906/416 | 12.9 | 2/68 | $<10^{-4}$ |
| | Median | 19237/592 | 12.7 | 2/67 | $<10^{-4}$ |
| Suppl. Exp 1 (size control) | Arithmetic Mean | 28257/1071 | 9.5 | 3/118 | $<10^{-4}$ |
| | Harmonic Mean | 21103/783 | 10.2 | 3/119 | $<10^{-5}$ |
| | Median | 23263/1222 | 8.6 | 3/120 | $<10^{-4}$ |
| Suppl. Exp 2 (fixation control) | Arithmetic Mean | 17410/1150 | 8.2 | 3/84 | $<10^{-4}$ |
| | Harmonic Mean | 12832/970 | 7.1 | 3/85 | $<10^{-3}$ |
| | Median | 15135/1232 | 7.1 | 3/79 | $<10^{-3}$ |

**Table S1. ANOVAs on arithmetic means, harmonic means and medians demonstrating robustness of the statistics**


**Figure S1 |** Reaction times per picture averaged across experiment 1, the size control experiment, and the fixation control experiment. Graphs show the RT per picture. Asterisks denote significance as determined with post-hoc tests; *$p < .05$; **$p < .01$; ***$p < 10^{-3}$. Error bars denote SEM after subtraction of the participants´ overall mean RTs.

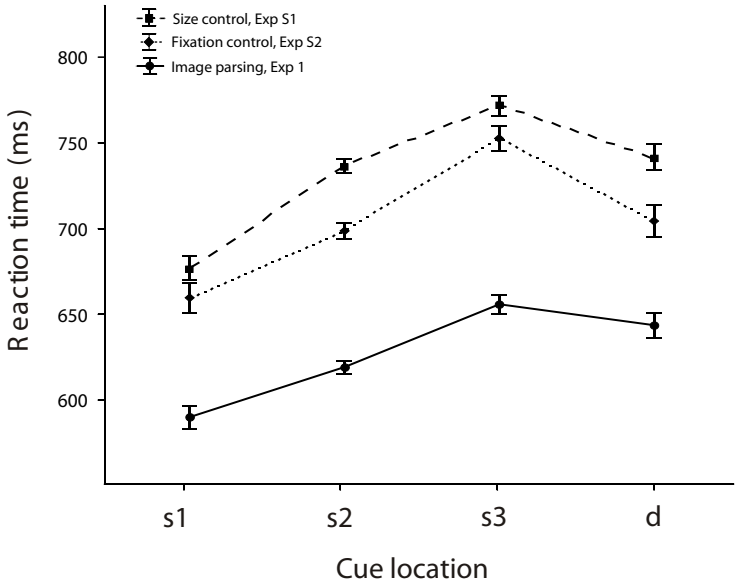


**Figure S2 |** Average reaction times in Experiment 1 (solid line) and the two control experiments (Size and Fixation). The dashed line shows RTs in the image-parsing task with the same stimuli presented at a smaller size. The dotted line shows RTs when eye-movements are restricted to central fixation. Restricting eye-moments and smaller pictures (with different response buttons) increased RTs but did not influence the slopes. The error bars denote SEM after subtraction of the participants' mean RT across all stimuli.