ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
2014, Vol. 25(5) 1087–1097
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797614522816
pss.sagepub.com
SAGE

Beyond Gist: Strategic and Incremental Information Accumulation for Scene Categorization

George L. Malcolm[1], Antje Nuthmann[2], and Philippe G. Schyns[3]
[1]The George Washington University, [2]University of Edinburgh, and [3]University of Glasgow

Abstract

Research on scene categorization generally concentrates on gist processing, particularly the speed and minimal features with which the "story" of a scene can be extracted. However, this focus has led to a paucity of research into how scenes are categorized at specific hierarchical levels (e.g., a scene could be a road or more specifically a highway); consequently, research has disregarded a potential diagnostically driven feedback process. We presented participants with scenes that were low-pass filtered so only their gist was revealed, while a gaze-contingent window provided the fovea with full-resolution details. By recording where in a scene participants fixated prior to making a basic- or subordinate-level judgment, we identified the scene information accrued when participants made either categorization. We observed a feedback process, dependent on categorization level, that systematically accrues sufficient and detailed diagnostic information from the same scene. Our results demonstrate that during scene processing, a diagnostically driven bidirectional interplay between top-down and bottom-up information facilitates relevant category processing.

Keywords
object recognition, eye movements

Received 11/30/12; Revision accepted 12/31/13

In order to interact with the world, people must quickly extract the semantic meaning of their environment to effectively guide their subsequent behaviors. Within around 100 ms of the onset of a scene, its "story," or gist, can be gleaned (Potter, 1975, 1976). These incredible scene-processing speeds have often led researchers to treat categorization as a perceptually driven process with semantic meaning passively emerging in a set hierarchy as critical properties become available to the visual system (Fei-Fei, Iyer, Koch, & Perona, 2007; Greene & Oliva, 2009a, 2009b; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Loschky & Larson, 2010; Oliva & Torralba, 2001; Rousselet, Joubert, & Fabre-Thorpe, 2005; Schyns & Oliva, 1994; Torralba & Oliva, 2003), perhaps as a result of a feed-forward network (Delorme, Richard, & Fabre-Thorpe, 2000; Serre, Oliva, & Poggio, 2007; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2002).

Consequently, most scene-recognition research has focused on the nature and extraction of the critical visual properties representing scene gist—that is, the minimal visual information enabling entry into a category hierarchy. Several candidates have emerged, such as the spatial regularities between volumetric forms (Biederman, 1987, 1995) or the variety of simple image statistics at low spatial resolutions (Greene & Oliva, 2009b; Oliva & Torralba, 2001, 2006; Schyns & Oliva, 1994). We contend that these perceptually driven accounts of gist have led the focus of scene-categorization research away from an important determinant of visual information: the categorization task itself (Oliva & Schyns, 1997; Rotshtein, Schofield, Funes, & Humphreys, 2010; Schyns, 1998).

Depending on a viewer's needs, a scene on a campus could be hierarchically categorized at the superordinate (e.g., indoor), basic (library), or subordinate (university library) levels (Gosselin & Schyns, 2001; Rosch, 1978;

Corresponding Author:
George L. Malcolm, Department of Psychology, The George Washington University, 2125 G St. NW, Washington, DC 20052
E-mail: glmalcolm@gwu.edu

Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Schyns, 1998; Tversky & Hemenway, 1983). The information accrued over time from the input scene could be constrained by these different categorization tasks, which suggests a bidirectional interplay between visual information in the input and information requirements in scene memory (Schyns, 1998).

The present study investigated whether there is evidence for this bidirectional interplay. We presented low-pass filtered scenes allowing just enough information for gist recognition. In addition, a gaze-contingent moving window (see McConkie & Rayner, 1975) enabled observers to freely supplement gist information with foveated full-resolution details. By recording fixation locations when basic and subordinate categorizations were performed, we identified not only the image coordinates of each fixation, but also modeled the specific full-resolution details and objects facilitating each judgment. If low-resolution image statistics (i.e., low spatial frequencies, or LSFs) suffice for gist processing, then full-resolution fixations are superfluous and should be randomly distributed over the input scene. Conversely, if access to the categorization hierarchy requires additional information, then observers will generate systematic fixations onto full-resolution diagnostic objects. Such evidence would go beyond gist by demonstrating the critical (and often neglected) feedback loop existing between the different information requirements of hierarchical scene categorizations and the incremental accrual of categorization-specific information from the same scene.

Method

Participants

Twenty-eight participants from the University of Edinburgh took part in the experiment. All reported normal or corrected-to-normal vision and gave informed consent before participating.

Apparatus

Stimuli were presented on a 21-in. ViewSonic (London, England) G225f CRT monitor (refresh rate = 140 Hz) positioned 90 cm from participants (25.18° × 19.01°). Participants sat with their head in a chin rest and made responses using a button box. We recorded eye movements using an SR Research (Mississauga, Ontario, Canada) EyeLink binocular desktop-mount system, equipped with a 2000-Hz camera upgrade, which allowed for binocular recordings at a sampling rate of 1000 Hz for each eye. Participants viewed stimuli binocularly, with the position of the gaze-contingent moving window determined by taking the mean of the two eye positions.

Stimuli and design

We selected 32 scenes from Google Images and converted them to gray scale (800 × 600 pixel resolution). Each scene belonged to one of four basic-level categories (restaurant, classroom, road, and pool), and within each basic category, each scene belonged to one of four subordinate categories (restaurant: diner, cafeteria, fine-dining establishment, pub; classroom: computer lab, lecture hall, preschool, elementary school; road: motorway, cul-de-sac, city street, roundabout; pool: Olympic pool, above-ground pool, Jacuzzi, water park). Thus, each scene could be categorized at two levels of specificity (Fig. 1).

We low-pass filtered scenes (i.e., removed high-spatial-frequency, or HSF, details) to leave only information below 25 × 18.75 cycles per image (0.83 × 0.62 cycles per degree of visual angle). A pilot test found that with this filtering, participants made basic and subordinate category judgments with 78% and 23% accuracy, respectively, during a four-alternative forced-choice (4AFC) task. In addition to showing participants low-pass-filtered images, we provided them with the ability to obtain additional visual details to inform their category decision: A gaze-contingent moving window, 2.5° in diameter, provided full-resolution information from the scene to foveal vision. Updating the display required 1 ms to receive gaze-position information from the eye tracker, less than 1 ms to draw the image textures, and up to 7 ms to refresh the screen. We smoothed the perimeter of the window with a Gaussian low-pass filter to avoid possible perceptual problems resulting from sharp-boundary windows (Fig. 2). The gaze-contingent window afforded viewers the ability to flesh out aspects of the scene with extra HSF details to facilitate category judgments, which would expose the feedback process.

Procedure

We assigned participants evenly to the basic and subordinate conditions.

Basic condition. In the basic condition, we first showed participants unfiltered examples from each of the four basic scene categories (none of these examples appeared in the experiment). When participants indicated that they understood the category labels, they then proceeded to the gaze-contingent experiment. The experiment started with four filler trials, with one scene from each basic category, followed by the 32 test scenes presented in random order. In each trial, the gaze-contingent window did not appear until after the first fixation, so if the information contained in the low-pass-filtered image was enough, then no eye movements would be

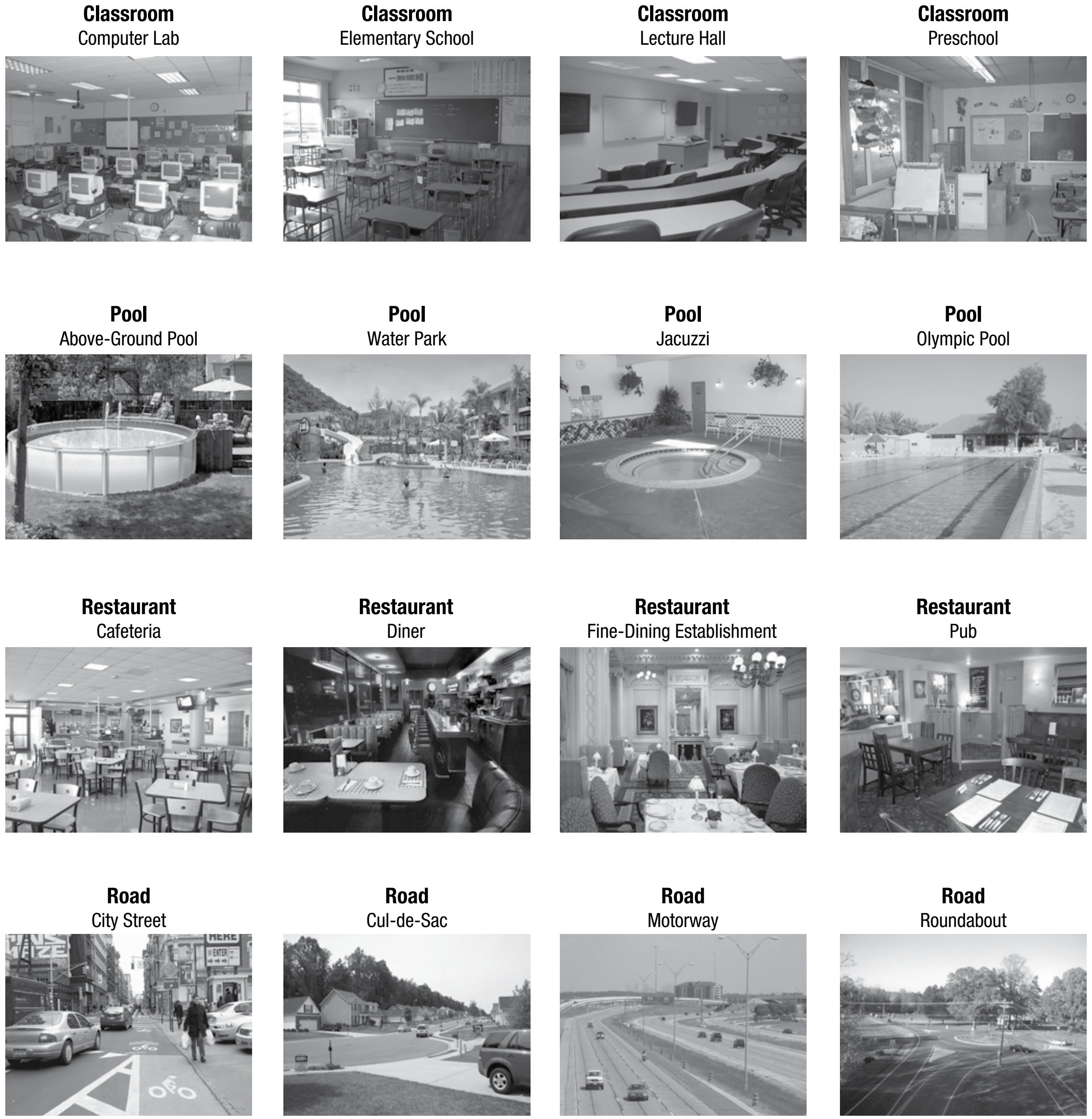


Fig. 1. Examples of scenes from each category. Each scene could be categorized into one of four groups, both at the basic level (shown in boldface) and the subordinate level (shown in regular type). Stimuli consisted of 32 scenes, half of which are shown here.

necessary. Participants viewed each scene until they were ready to make a category response, which they did by pressing a button. The scene then disappeared, and participants chose which category the scene belonged to using a separate set of buttons. Trials timed out after 15 s if the button was not pressed.

Subordinate condition. The paradigm in the subordinate condition was the same as in the basic condition, with the following exceptions. We divided test scenes into blocks by their basic-level category. Each block now started with an unfiltered example from each subordinate category. Once participants indicated that they understood

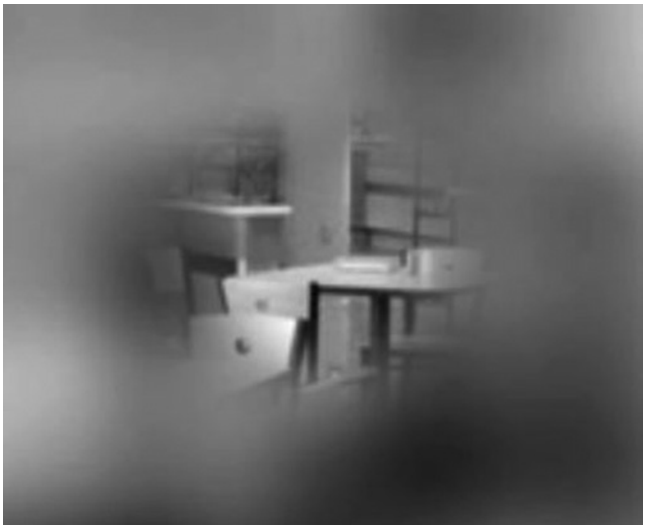


Fig. 2. Example of what a participant might have seen during the gaze-contingent experiment. Each scene was presented at low spatial frequencies. However, an area 2.5° in diameter appeared in full resolution wherever participants moved their eyes. A Gaussian low-pass filter was applied around the border of the window to avoid perceptual problems resulting from sharp boundaries.

the category labels, the block of trials began with four filler trials, one from each subordinate category within the basic group, followed by the eight test scenes in random order. Blocks appeared in random order.

Results

We excluded incorrect trials in which the wrong category was selected (16 and 27 from the basic and subordinate conditions, respectively) and trials that timed out after 15 s (2 from the subordinate group). The remaining 432 and 419 trials eligible for analysis comprised 96.4% and 93.5% of the basic and subordinate trials, respectively. Analyses also took into account which basic-level category (henceforth called stimuli type) the scene belonged to, thereby allowing for a 2 × 4 mixed-design analysis of variance (ANOVA) with judgment (basic vs. subordinate) as a between-subjects factor and stimuli type (classroom, pool, restaurant, road) as a within-subjects factor. Whenever assumption of sphericity was violated, we used a Greenhouse-Geisser correction.

We found a main effect of judgment on response time (RT); processing was faster in the basic condition (M = 2,305 ms) than in the subordinate condition (M = 3,253 ms), $F(1, 26) = 9.277$, $p = .005$, $\eta_p^2 = .263$. There was also a main effect of stimuli type, $F(3, 78) = 12.910$, $p < .001$, $\eta_p^2 = .332$ (*M*s = 2,832, 2,709, 3,419, and 2,157 ms for classrooms, pools, restaurants, and roads, respectively), but no interaction between the two factors, $F(3, 78) = 1.805$, $p = .153$, $\eta_p^2 = .065$ (Table 1).

The RT difference suggests that participants' ability to categorize a scene depends on the information needed for each level of specificity. However, it is not clear what participants were doing during this time: If participants in the subordinate condition used the extra time to accrue HSF information to facilitate their judgment, they should have made more eye movements than participants in the basic condition. Conversely, if participants found all their needed information in the low-pass-filtered image, and the increased RT was simply due to a longer decision-making process, then there should be a similar number of fixations across categorization conditions.

Table 1. Results From the Gaze-Contingent Experiment: Means for Key Variables for Each of the Four Scene Categories in Each Condition

	Classroom		Pool		Restaurant		Road	
Measure	Basic condition	Subordinate condition	Basic condition	Subordinate condition	Basic condition	Subordinate condition	Basic condition	Subordinate condition
Response time (ms)	2,402 (252)	3,262 (352)	2,428 (269)	2,990 (245)	2,676 (212)	4,162 (366)	1,715 (170)	2,598 (331)
Number of fixations	5.74 (0.72)	9.77 (1.12)	5.64 (0.60)	8.82 (0.93)	6.54 (0.64)	12.17 (1.27)	3.66 (0.32)	8.11 (1.27)
Initial saccade latency (ms)	419 (26.61)	371 (22.97)	418 (32.26)	327 (16.98)	437 (37.34)	357 (19.09)	411 (25.33)	351 (17.61)
Fixation duration (ms)	369 (21.14)	290 (18.93)	383 (24.07)	325 (27.00)	377 (22.80)	306 (19.40)	394 (23.17)	277 (15.47)
Saccade amplitude (degrees)	2.55 (0.25)	2.96 (0.24)	2.14 (0.17)	2.47 (0.24)	2.70 (0.24)	2.73 (0.26)	2.50 (0.28)	2.52 (0.25)

Note: Standard errors are given in parentheses. Number of fixations did not include the initial fixation at scene onset. Fixation durations did not include the initial saccade latency.

We found a main effect of judgment on fixation count, with more fixations made during subordinate categorizations (M = 9.72) than during basic categorizations (M = 5.40), F(1, 26) = 17.906, p < .001, η_p^2 = .408. There was also a main effect of stimuli type, F(3, 78) = 9.558, p < .001, η_p^2 = .269 (Ms = 7.76, 7.23, 9.36, and 5.89 for classrooms, pools, restaurants, and roads, respectively), but no interaction, F(3, 78) = 1.201, p = .315, η_p^2 = .044 (Table 1).

Thus, participants did indeed generate eye movements to supplement visible LSF information with HSF details. Such details were acquired even during basic-level categorization, which demonstrates that viewers use HSF information when available. Additionally, participants in the subordinate condition, compared with those in the basic condition, needed more time and fixations—this suggests that performing subordinate-level categorizations is a separate, slower information-accrual process than performing basic-level categorizations.

Eye movement strategies

These results suggest that scene categorization demands different information depending on the required level of specificity. In order to further probe the changing nature of information accrual across hierarchical levels, we examined respective eye movement strategies. Both basic and subordinate categorization required HSF information in addition to LSF information, but if the respective sampling processes differed, then eye movement strategies should have differed as well.

We first examined initial saccade latency: the time between scene onset and the initiation of the first saccade. There was a main effect of judgment, F(1, 26) = 5.720, p = .024, η_p^2 = .180, with basic categorizations having longer saccade latencies (M = 421 ms) than subordinate categorizations (M = 352 ms), but no effect of stimuli type or interaction with judgment (Fs < 1; Table 1). Thus, even within the first approximately 350 ms of scene onset, when the exact same visual image was available (the full-resolution window did not appear until after the first saccade), the respective information demands affected category processing.

We then examined fixation durations (not including the initial saccade latency) and saccade amplitudes. There was a main effect of judgment on fixation duration, F(1, 26) = 8.246, p = .008, η_p^2 = .241 (basic: M = 381 ms; subordinate: M = 300 ms), a trend for stimuli type, F(3, 78) = 2.242, p = .09, η_p^2 = .079 (Ms = 329, 354, 341, and 335 ms for classrooms, pools, restaurants, and roads, respectively), as well as a significant interaction, F(3, 78) = 3.207, p = .028, η_p^2 = .110. This interaction occurred because participants in the basic condition had numerically longer fixations when categorizing pool scenes (M = 383 ms) compared with participants in the subordinate condition (M = 325 ms), t(26) = 1.598, p = .122; fixation durations for all other scenes were significant, ts > 2.370, ps < .025 (Table 1). There was no effect of judgment on saccade amplitudes (F < 1); however, as one would expect given the different layouts between the different category groups, there was a main effect of stimuli type, F(3, 78) = 5.882, p = .001, η_p^2 = .184 (Ms = 2.75°, 2.31°, 2.72°, and 2.51° for classrooms, pools, restaurants, and roads, respectively; Table 1). There was no interaction, F(3, 78) = 1.444, p = .237, η_p^2 = .053. These results demonstrate that participants' initial and subsequent fixation durations differed depending on the categorization being made, a finding that supports different ongoing processes.

Gaze patterns

Within the categorization hierarchy, there are more common attributes at the subordinate than at the basic category level; however, there are generally more features that allow viewers to differentiate one basic category from another than there are that allow viewers to differentiate one subordinate category from another (Rosch et al., 1976). Thus, subordinate diagnostic information is comparatively sparser and more difficult to locate in the image than the associated basic diagnostic information. This difference should affect information-sampling strategies. We analyzed gaze patterns over time by measuring the distance of fixations from the center of the image over the first five fixations. We found fixations distributed in a center-surround pattern for basic and subordinate judgments, respectively (Fig. 3).

A three-way mixed-design 2 (judgment: basic, subordinate) × 4 (stimuli type: classroom, pool, restaurant, road) × 5 (ordinal fixation: 1, 2, 3, 4, 5) ANOVA on fixation distance from screen center revealed a critical interaction between judgment and ordinal fixation, F(4, 88) = 3.661, p = .008, η_p^2 = .143; this interaction indicates that participants carried out different sampling strategies over the first five fixations depending on the judgment they were assigned to make (Fig. 4).

Participants directed fixations further into the periphery of the scene during their first five fixations in the subordinate condition than in the basic condition, which suggests a wider search for diagnostic object information. For the sake of completeness, there was no main effect of judgment, F(1, 22) = 2.163, p = .156, η_p^2 = .090, though there was an effect of stimuli type, F(3, 66) = 4.342, p = .007, η_p^2 = .165 (Ms = 3.88°, 3.47°, 3.87°, and 3.52° for classrooms, pools, restaurants, and roads, respectively) as well as of ordinal fixation, F(2.286, 50.288) = 42.240, p < .001, η_p^2 = .658 (Ms = 2.55°, 3.28°, 3.86°, 4.26°, and 4.48° from the center for Fixations 1 through 5, respectively). There was no interaction between stimuli type and ordinal count (F < 1), no interaction between judgment and

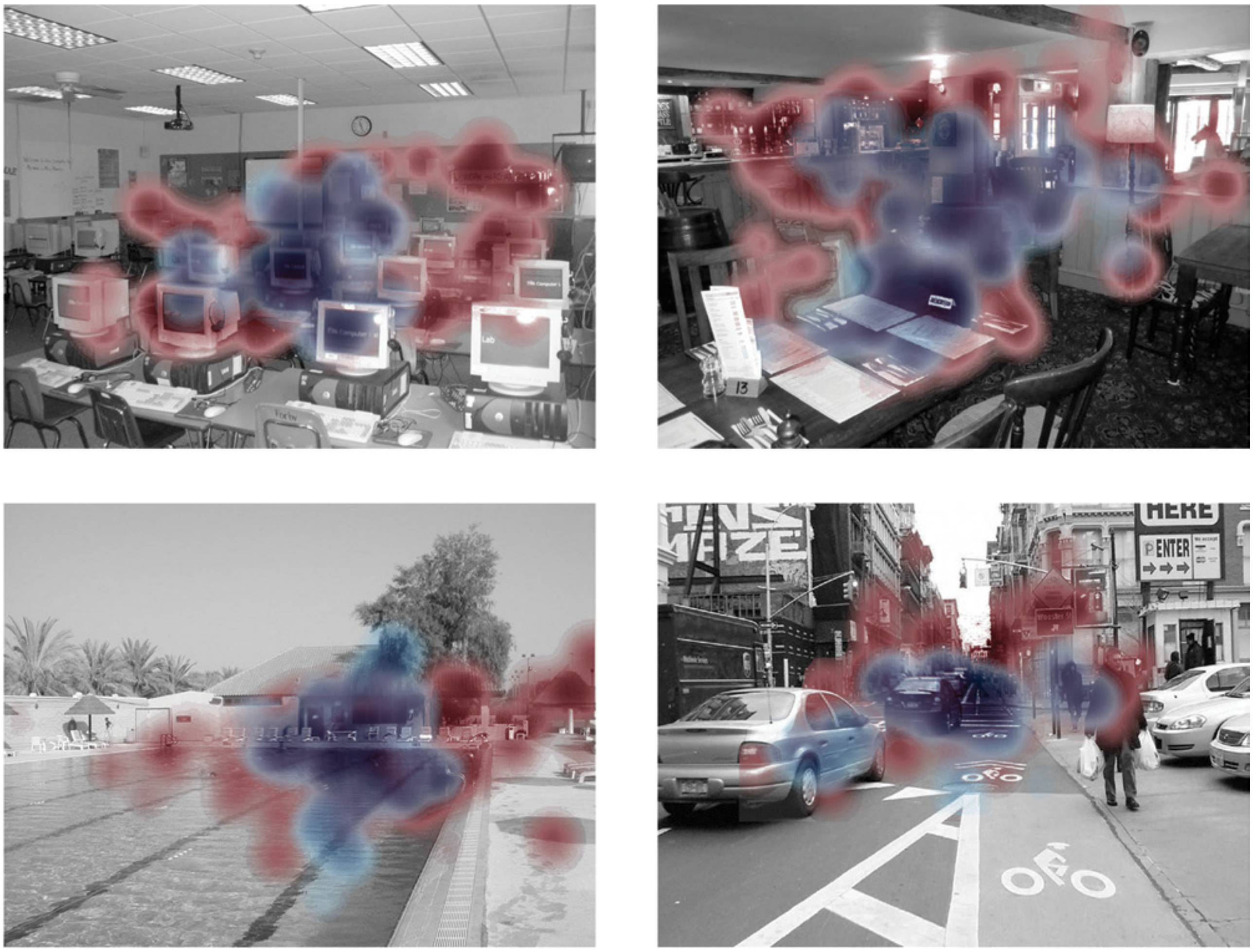


Fig. 3. All fixations made by participants in the basic condition (blue) and subordinate condition (red), during the gaze-contingent experiment. Overlapping areas appear as purple. Results are shown for one scene from each category.

stimuli type, $F(3, 66) = 2.012$, $p = .121$, $\eta_p^2 = .084$, and no three-way interaction ($F < 1$).

Diagnostic objects

To ascertain which objects were diagnostic during categorization at each level, we calculated the mean proportion of fixations on each object in both conditions. All objects that fell within the gaze-contingent window during a fixation were considered fixated. To determine statistical validity, we compared the mean proportion of fixations per object and condition against the overall fixation proportion in a bootstrap analysis using 14,000 repetitions.

In both the basic and subordinate conditions, each scene contained significantly fixated objects ($ps < .05$), which indicates that participants always accrued specific object information to complement the LSF gist. Additionally, as every scene contained at least one object—and usually many more—that was diagnostic for only one level, it is clear that participants strategically directed gaze to find and extract hierarchy-specific diagnostic information (Fig. 5).

Validation of diagnostic objects

To verify that these significant objects were diagnostic, we asked 36 new participants to categorize scenes at either the basic or subordinate level using the same paradigm as in the gaze-contingent experiment, except that scenes were now masked after 150 ms, and eye movements were not recorded. During presentation, each scene was low-pass filtered except for the objects previously identified as diagnostic at either the basic or

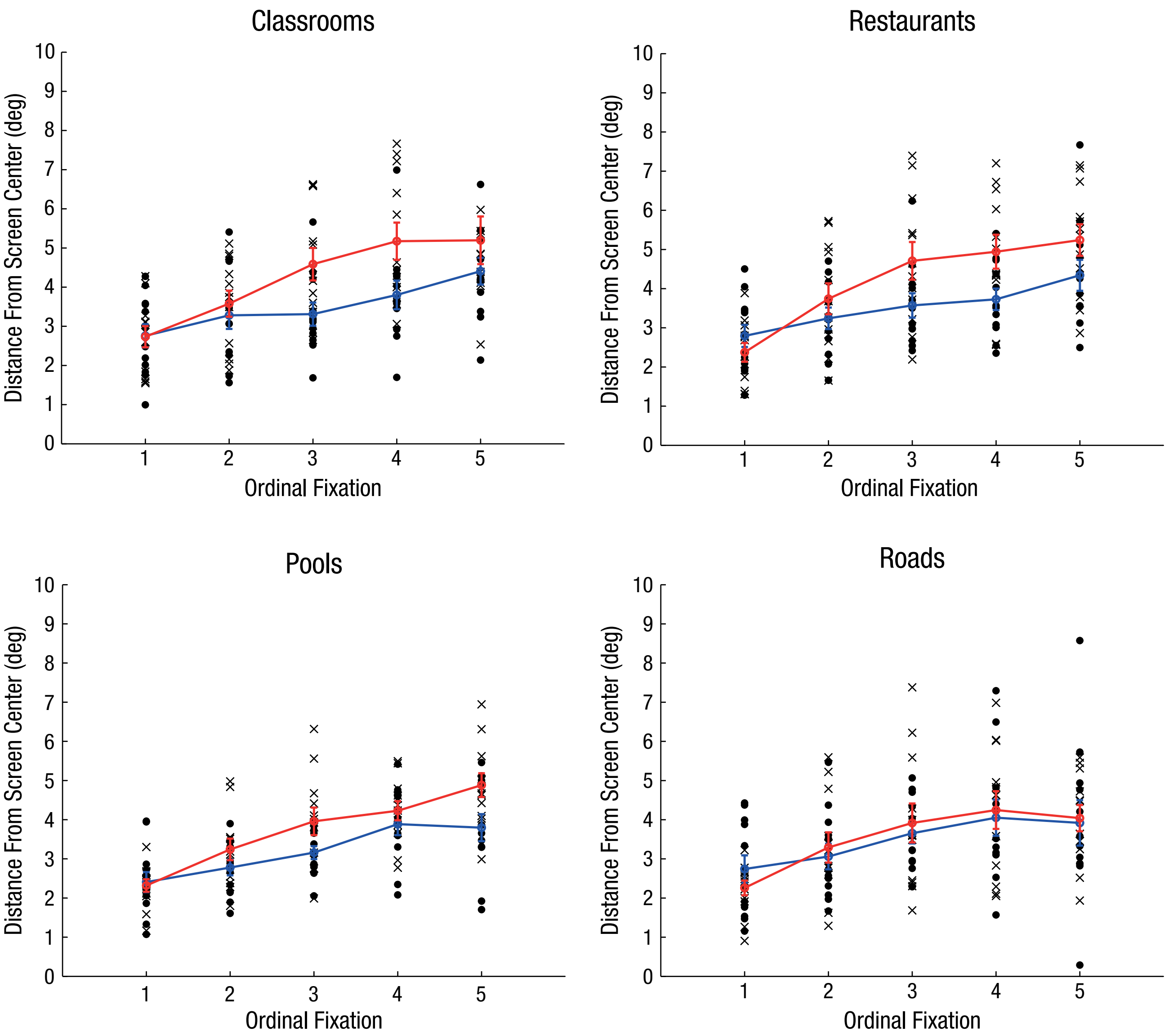
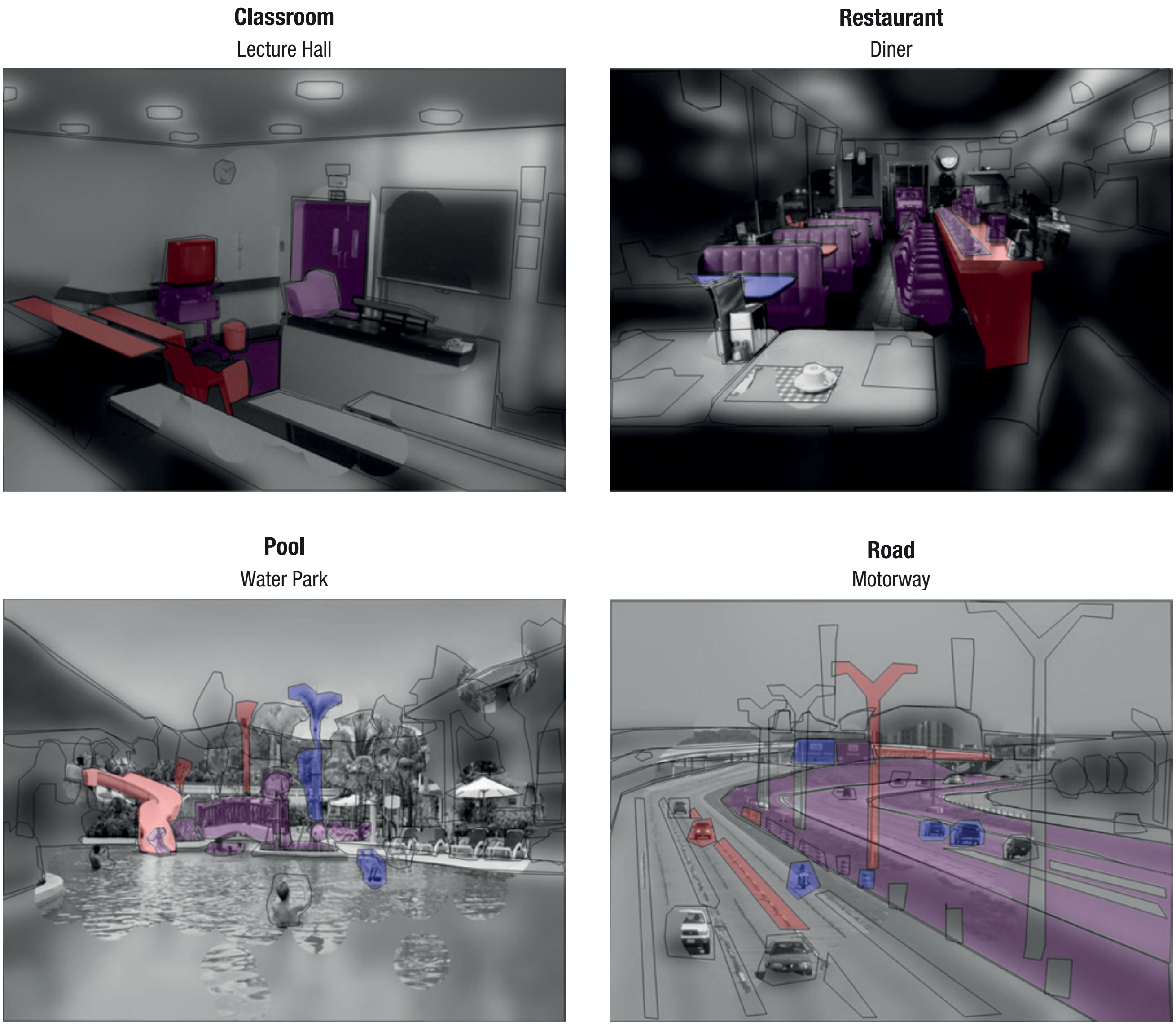


Fig. 4. Distance of fixations from the center of the screen as a function of ordinal fixation, separately for the four basic categories of stimuli. Black circles and crosses represent individual means for each participant in the basic and subordinate conditions, respectively. Means across the basic and subordinate conditions are indicated, respectively, by blue and red circles (lines connecting the condition means are shown for clarity). Errors bars represent ±1 *SE*.

subordinate level, which were shown without filtering (Fig. 6). The two types of diagnostic information (basic vs. subordinate) for each scene were distributed evenly across participants in both categorization groups, creating a 2 × 2 mixed design (because there were only eight scenes per basic category and we were measuring accuracy, stimuli type was collapsed). We predicted that accuracy should be high when making a categorization with diagnostic information at the same level (i.e., subordinate objects with a subordinate judgment and basic objects with a basic judgment). We further predicted that because comparatively fewer objects distinguish one subordinate category from another than distinguish one basic category from another (Rosch et al., 1976), swapping the diagnostic objects (i.e., presenting basic objects for a subordinate judgment and vice versa) should have a minimal effect on basic-level categorization but a larger effect on subordinate judgments.

We next calculated the mean proportion of hits in both conditions. Analyses revealed a main effect of judgment on accuracy, $F(1, 34) = 9.339$, $p = .004$, $\eta_p^2 = .215$ (basic: $M = .79$; subordinate: $M = .71$) and a main effect



Classroom

Lecture Hall

Restaurant

Diner

Pool

Water Park

Road

Motorway

Fig. 5. Illustration showing fixated scene regions in the basic and subordinate conditions. Each panel shows a low-pass-filtered version of a scene from one of the four categories in which every fixated area in both categorization groups is mapped in full resolution (objects are outlined here for clarity). Blue regions indicate objects significantly fixated in the basic condition, red regions indicate objects significantly fixated in the subordinate condition, and purple areas indicate objects significantly fixated in both conditions.

of diagnostic information, $F(1, 34) = 15.871$, $p < .001$, $\eta_p^2 = .318$ (basic: $M = .71$; subordinate: $M = .79$). There was also a trend for an interaction, $F(1, 34) = 3.456$, $p = .072$, $\eta_p^2 = .092$ (Fig. 6), which was due to a significant change in subordinate categorization accuracy when basic or subordinate diagnostic objects were visible (Ms = .65 and .76, respectively), $t(17) = 4.134$, $p = .001$, but no significant difference during basic categorization (basic diagnostic objects: $M = .77$ and subordinate diagnostic objects: $M = .81$), $t(17) = 1.506$, $p = .151$. The results confirm that the objects fixated during the gaze-contingent experiment were indeed diagnostic for the respective categorization processes.

Behavior verification

To investigate whether categorizing normal, full-resolution scenes involves similar sampling strategies, we asked 28 new participants to categorize full-resolution images with no gaze-contingent window. Half were assigned to make basic judgments, and half were assigned to make subordinate judgments. During these normal viewing conditions, there was a strong trend of judgment on RT, $F(1, 26) = 3.568$, $p = .070$, $\eta_p^2 = .121$, with basic categorization needing less time (1,062 ms) than subordinate categorization (1,272 ms). There was a main effect of stimuli type, $F(2.084, 54.188) = 7.080$, $p = .002$, $\eta_p^2 = .214$ (Ms = 1,269,

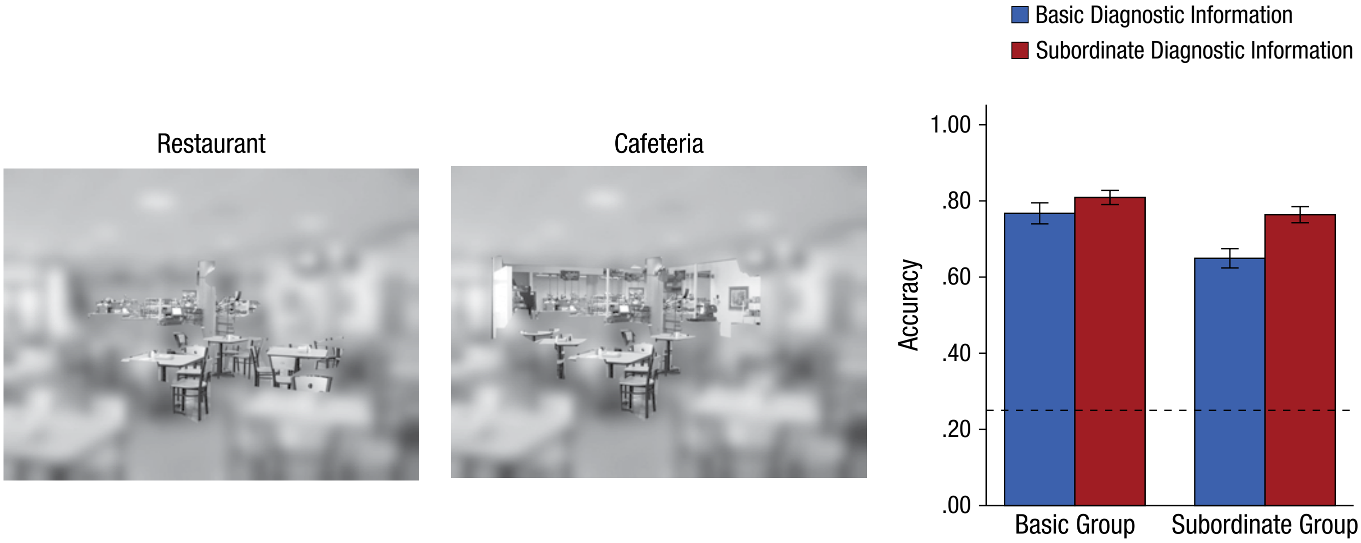


Fig. 6. Paradigm and results from the validation experiment. Participants were asked to categorize individual scenes, but unlike in the gaze-contingent experiment, each scene contained objects identified as diagnostic at either the basic level (example shown on left) or subordinate level (example shown in middle). These objects were revealed at full resolution while the rest of the scene was low-pass filtered. The graph shows mean accuracy as a function of condition and the type of diagnostic information presented. The dashed line indicates chance performance. Error bars represent ±1 *SE*.

1,128, 1,273, and 998 ms for classrooms, pools, restaurants, and roads, respectively) as well as a trend for an interaction, $F(3, 78) = 2.596$, $p = .058$, $\eta_p^2 = .091$, which was due to restaurants taking significantly longer to categorize at the subordinate level (M = 1,471 ms) than at the basic level (M = 1,075 ms), $t(26) = 3.00$, $p = .006$. Similar numerical patterns were obtained for classrooms (basic: M = 1,204 ms; subordinate: M = 1,333 ms), pools (basic: M = 989 ms; subordinate: M = 1,266 ms), and roads (basic: M = 979 ms; subordinate: M = 1,016 ms), but they were not significant (ts < 1.629, ps > .115).

Fixation counts showed a similar pattern of results, with basic categorization requiring fewer fixations (M = 2.95) than subordinate categorization (M = 4.01), $F(1, 26) = 5.391$, $p = .028$, $\eta_p^2 = .172$. There was also a main effect of stimuli type, $F(1.909, 49.641) = 8.239$, $p = .001$, $\eta_p^2 = .241$ (Ms = 3.96, 3.21, 3.85, and 2.89 for classrooms, pools, restaurants, and roads, respectively), and a trend for an interaction, $F(3, 78) = 2.637$, $p = .055$, $\eta_p^2 = .092$, which was due to significantly more fixations during subordinate than during basic decisions for pools (basic: M = 2.52; subordinate: M = 3.91), $t(26) = 1.901$, $p = .068$, and for restaurants (basic: M = 3.01; subordinate: M = 4.69), $t(26) = 3.363$, $p = .002$. In contrast, there was a similar numerical pattern but no significant difference for classrooms (basic: M = 3.57; subordinate: M = 4.34) and roads (basic: M = 2.69; subordinate: M = 3.09), ts < 1.415, ps > .169. The similarity of the RT and fixation-count results with those from the gaze-contingent experiment demonstrates that even when viewing full-resolution images, participants need more time and fixations to accrue diagnostic information when making subordinate decisions.

Additional analyses examined eye movement behaviors, including initial saccade latency, fixation duration, and saccade amplitude. For initial saccade latency, there was no main effect of judgment, $F(1, 26) = 1.708$, $p = .203$, $\eta_p^2 = .062$, or stimuli type, nor was there an interaction between the two for initial saccade latency (Fs < 1). Fixation durations showed no main effect of judgment (F < 1), though there was an effect of stimuli type, $F(2.288, 59.487) = 4.539$, $p = .011$, $\eta_p^2 = .149$, (Ms = 257, 312, 265, and 277 ms for classrooms, pools, restaurants, and roads, respectively). There was no interaction, $F(3, 78) = 1.276$, $p = .288$, $\eta_p^2 = .047$. No main effect of saccade amplitude on judgment appeared, $F(1, 26) = 1.122$, $p = .299$, $\eta_p^2 = .041$, but there was an effect of stimuli type, $F(3, 78) = 20.836$, $p < .001$, $\eta_p^2 = .445$ (Ms = 4.06°, 3.33°, 4.31°, and 3.44° for classrooms, pools, restaurants, and roads, respectively), and an interaction, $F(3, 78) = 4.869$, $p = .004$, $\eta_p^2 = .158$, with restaurants having a trend for longer saccades during subordinate categorizations (M = 4.62°) than during basic categorizations (M = 3.99°), $t(26) = 1.735$, $p = .094$. In contrast, there were similar, but not significant, numerical patterns for classrooms (basic: M = 3.81°; subordinate: M = 4.31°) and pools (basic: M = 3.04°; subordinate: M = 3.63°), ts < 1.560, ps > .131, whereas for roads, there were longer saccades in the basic condition (M = 3.61°) than in the subordinate condition (M = 3.27°), t < 1.

In sum, the consistent RT and fixation count results between the gaze-contingent experiment and the unfiltered-image control experiment suggest that even when participants view unfiltered images, subordinate categorizations are a separate, slower process than basic categorizations. However, fixation durations are no longer affected as a function of task when participants view full-resolution images.

General Discussion

Gist-processing research has generally focused on identifying the minimal common attributes that distinguish categories rather than examining the information demands that bias the categorization process. Here, we demonstrated that during scene viewing, there is a feedback component that creates a bidirectional interplay between available information and task goals and that this leads to categorization at a specified level.

In particular, our data reveal that LSF information does not provide the full range of diagnostic information for basic-level access: The several fixations made prior to a category judgment indicate that there is often a need for additional object processing. Our data indicate that observers go beyond assumed LSF gist information to find an entry point into the categorical hierarchy, which means that global, coarse scene representations might be necessary but are not always sufficient. Given the nature of the paradigm we used, it is possible that some of the fixations were made in order to verify that a basic category decision was correct. However, even in such situations, coarse global properties clearly did not provide enough information to rapidly process gist, so additional object information was needed. Determining gist can therefore be thought of as a process of information sampling across multiple scales of information.

In addition, the sampling strategy during basic categorization was found to differ from the strategy used during subordinate categorizations, which indicates that strategic acquisition processes vary as a function of the task and further corroborates that information demands bias the scene-categorization process. Because the original LSF information at scene onset was the same across both conditions, this difference in sampling extends soft-wired hypotheses (Oliva & Schyns, 1997) because it suggests that the visual system does not rely on a single mode of processing but that given the range of information available, it can adjust a sampling strategy to categorize an image with maximum efficiency. It should be noted that participants in our subordinate condition already knew the basic-level category prior to all trials, which possibly influenced their sampling strategy. However, given that the goal was to isolate the sampling strategy at the basic and subordinate levels, this detail was necessary to separate the two processes.

Now that we have uncovered sampling strategy markers for these two hierarchical levels, it would be interesting to investigate with more ecologically valid, open-ended categorizations (e.g., participants categorize scenes at the subordinate level with no a priori basic-level knowledge) when the shift in strategies takes place. It could be a strictly sequential process in which basic-level categorization must be completed prior to subordinate-related information sampling, or there could be an overlap in which subordinate sampling strategies begin while basic-level categorizations are not fully formed. Such a distinction would help refine the understanding of how the viewer samples and integrates diagnostic information prior to a judgment.

Author Contributions

G. L. Malcolm and P. G. Schyns developed the study concept. All authors contributed to the study design. Testing and data collection were performed by G. L. Malcolm and A. Nuthmann. G. L. Malcolm analyzed the data, and all authors interpreted it. G. L. Malcolm drafted the manuscript, and A. Nuthmann and P. G. Schyns provided critical revisions. All authors approved the final version of the manuscript for submission.

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

G. L. Malcolm was supported by Economic and Social Research Council Grant PTA-026-27-2500.

References

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

Biederman, I. (1995). Visual object recognition. In S. F. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science* (2nd ed., Vol. 2, pp. 121–165). Cambridge, MA: MIT Press.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research*, *40*, 2187–2200.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), Article 10. Retrieved from http://www.journalofvision.org/content/7/1/10

Gosselin, F., & Schyns, P. G. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological Review*, *108*, 735–758.

Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*, 464–472.

Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–179.

Joubert, O., Rousselet, G., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*, 3286–3297.

Loschky, L. C., & Larson, A. M. (2010). The natural/man-made distinction is made prior to basic-level distinctions in scene gist processing. *Visual Cognition*, *18*, 513–536.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, *17*, 578–586.

Oliva, A., & Schyns, P. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*, 72–107.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*, 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research: Visual Perception*, *155*, 23–36.

Potter, M. C. (1975). Meaning in visual scenes. *Science*, *187*, 965–966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 509–522.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 28–48). Hillsdale, NJ: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rotshtein, P., Schofield, A., Funes, M. J., & Humphreys, G. W. (2010). Effects of spatial frequency bands on perceptual decision: It is not the stimuli but the comparison. *Journal of Vision*, *10*(10), Article 25. Retrieved from http://www.journalofvision.org/content/10/10/25.full

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes? *Visual Cognition*, *12*, 852–877.

Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition*, *67*, 147–179.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*, 195–200.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, USA*, *104*, 6424–6429.

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

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, *15*, 121–149.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, *42*, 2593–2615.