Human Learning of Contextual Priors for Object Search: Where does the time go?

Barbara Hidalgo-Sotelo[1] Aude Oliva[1] Antonio Torralba[2]

[1]Department of Brain and Cognitive Sciences,
[2]Computer Science and Artificial Intelligence Laboratory
MIT, Cambridge, MA, 02139
bhs@mit.edu, oliva@mit.edu

Abstract

Attention allocation in visual search is known to be influenced by low-level image features, visual scene context and top down task constraints. Here, we investigate the role of Contextual priors in guiding visual search by monitoring eye movements as participants search very familiar scenes for a target object. The goal of the study is to identify which stage of the visual search benefits from contextual priors. Two groups of participants differed in the expectation of target presence associated with a scene. Stronger priors are established when a scene exemplar is always associated with the presence of the target than when the scene is periodically observed with and without the target. In both cases, overall search performance improves over repeated presentations of scenes. An analytic decomposition of the time course of the effect of contextual priors shows a time benefit to the exploration stage of search (scan time) and a decrease in gaze duration on the target. The strength of the contextual relationship modulates the magnitude of gaze duration gain, while the scan time gain constitutes one half of the overall search performance benefit regardless of the probability (50% or 100%) of target presence. These data are discussed in terms of the implications of context-dependent scene processing and its putative role in various stages of visual search.

1. Contextual Priors in Object Search

Experience, in a broad sense, may be regarded as a means of biasing a set of expectations based on perceived situational regularities [8]. Goal-directed action, such as visual search, represents an ecologically relevant facet of experience, as repeatedly performing a search in a similar environment can often influence search performance [2, 3, 9, 17]. The search of my cluttered office for a cell phone, for example, is abbreviated by previously established knowledge that my phone is most likely located next to the lamp on my desk. Attentional emphasis on spatial locations with a high probability of target presence may thus improve search performance by drawing on established knowledge about a scene and target position.

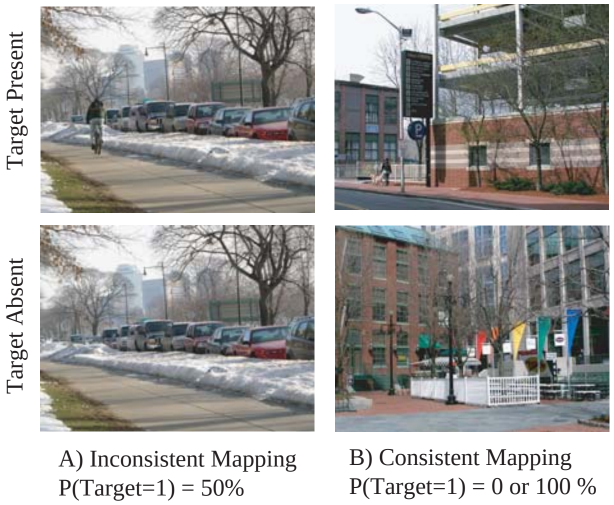


Figure 1: Illustration of the two experimental conditions of identity priors: Inconsistent mapping where the association between scene and target presence is 0.5 and consistent mapping where the association between a particular scene and the presence of the target is either 1 either 0.

The association between my phone and its typical place on my desk represents a specific instantiation of context designated an *Identity prior*. Varied strengths of identity priors are constructed from experience and yield a set of expected target locations. Even when visiting a new office, however, performing a search for “phone” will exhibit a common pattern of scanning waist-level surfaces. *Categorical priors* constitute one form of top-down contextual guidance and are triggered by the activation of semantic knowledge between an object and a scene category.

When looking for an object in a real-world scene, attention and eye movements are guided towards informative regions likely to contain the target object [5]. Categorical contextual priors exert their effects on search in a pre-attentive manner, as evidenced by findings from Torralba and collaborators [16, 10] indicating that more than 80% of first eye fixations land within the categorically-specified contextual region. Informative regions may be determined, not only by categorical context, but by individual experience as well. Which stages of visual processing and task constraints are susceptible to the influence of contextual identity priors?

*Identity prior*s provide a framework in which to consider the association between a specific scene exemplar and the location of objects. Contextual guidance by identity priors activates a strong predictive relationship between types of objects and their configurations in a given scene. Performing natural tasks yields fixation patterns that are heavily modulated by specific task demands and local visual context [4, 13]. Everyday activities involving sequential actions, food preparation for example, result in fixations of predominantly task-relevant objects; eye movements, as such, are controlled by primarily ”top down” information [7]. In contextual cuing, repeated exposure to global pattern information can produce implicit learning of the layout of the scene sufficient to guide attention and help target detection [2, 3, 11].

To what extent does the strength of the contextual relationship influence the timing of this contribution? At what stage of visual processing do strong identity priors contribute to scene or object recognition: at an early stage (i.e. before the first eye movement is initiated), during the exploration stage (i.e. the visual search per se), and/or at a later, decision (i.e. target processing) stage?

Two experimental conditions were designed to model different strengths of identity priors. The results of the experiment describe the effects of identity priors on the magnitude of search improvement over repeated search of familiar scenes, as well as a difference in the susceptibility of visual search stages to guidance by identity priors. In both conditions an improvement in overall search performance is observed, with the magnitude of improvement depending on the strength of identity priors.

2. Behavioral Experiment

Two conditions (depicted in Figure 1) are designed to model varying weights of identity priors: consistent mapping and inconsistent mapping [12]. In both conditions, scene stimuli are shown repeatedly and in each case the search task is to respond whether the target- a person- was present in the scene. At the outset, the marginal probability of a target being present anywhere in a scene is 0.5, as half of the stimuli contained a person and the other half did not.

Over the course of many repetitions, conditional probabilities are established such that presentation of any given scene cues a particular location to be biased toward target presence or absence. In the inconsistent case (Figure 1A), identical versions of a scene are shown with equal frequency, while a target appears at a single location in half of the trials and is absent from the other half of the trials. In the consistent case (Figure 1B), a given scene will be either target-containing or target-lacking. Identity priors in this condition constitute the strongest relationship between a scene and target object status. Given a consistently-mapped scene, the probability of a target being located in the scene is thus either 0 or 1. We can therefore make two predictive observations.

First, once identity priors have been established scene recognition alone should be sufficient to diagnose target status in the consistent condition. In theory no eye movements should be necessary to determine the location or presence of a target, if the scene priors are used prior to programming the first saccade. Secondly, recognition of a scene in the inconsistent condition should serve only as a cue of likely target location, while an accurate determination requires that the gaze approach the target region.

We expect that learning identity priors will facilitate search as a specific contextual relationship is established between local target and global scene context. It is expected that overall search performance (measured as reaction time in pressing the target present key) will be more efficient in consistent mapping of a one scene to a target and a different scene to no target. Initially, all scenes are novel and no difference should be observed between the two conditions. Over many scene exposures we expect search efficiency to reflect the building of identity priors, which are stronger when mapping is consistent. Even in inconsistent mapping, however, we believe that the expectation of a target at one location will rapidly adjust attentional allocation such that search is improved over the course of many trials. Depending on the temporal characteristics of identity prior activation, the improvement in search may reflect the susceptibility of various stages of search to guidance by varying degrees of contextual information.

2.1. Apparatus

An ETL 400 ISCAN table-mounted, video-based eye tracking system was used to record eye position during the course of the experiment. This system samples eye position at a rate of 240 Hz and is accurate to within 0.5 degrees of the visual angle. Participants are seated 75 cm from the presentation screen, 65 cm from the eye tracking camera, and adjust a headrest to comfort. At these specifications, the eye tracker covers a visual angle of 30 x 20 degrees. The camera is aligned with the right eye of the subject; contrast, zoom, and focus are manually adjusted to allow full visibility of the eye when looking at any part of the screen. Thresholds for tracking the corneal and pupil reflections are adjusted manually to maximize tracking efficiency for each subject. A five point calibration is used, during which the coordinates of the pupil and corneal reflection are recorded for positions in the center and each corner of the screen. During the experiment, position data is transmitted from the eye tracking computer to the presentation computer so as to ensure fixation of a cross in the center of the screen for 500 ms prior to stimulus presentation.

2.2. Procedure

Thirty-two participants (sixteen in each condition) were recruited and paid $10 to participate in the experiment. Participants were consenting 18-40 years old with normal or contact lens-corrected vision. Forty-eight scene stimuli, each existing in two versions (one target present, one target absent), were randomized and counterbalanced across groups and conditions. The scenes were depicted by photographs of park and city environments. A pretest was conducted to establish that target presence could not be detected at a glance (200 ms). Participants were instructed to detect the presence (or absence) of a person in the image and press the corresponding key as soon as a determination was made; this search was performed over 20 successive blocks (comprised of 48 stimuli). In the inconsistent case, identical versions of a scene are shown with equal frequency, with a target appearing at a single location in half of the trials; the target is absent from the other half of the trials. In the consistent case, a given scene will be either target-containing or target-lacking. Eye movements were recorded throughout the course of the experiment. After each repetition, the calibration is checked using a visual assessment of tracking accuracy on the five point calibration screen. Every 5 blocks, the subject was given a break. Each subject performed a total of 960 trials, resulting in average experiment duration of 45 minutes.

2.3. Analyses

For each participant, we registered four measures: (1) the overall reaction time per image, (2) the initial (or central) fixation duration, (3) the scan time, from movement away from initial fixation until the first entry into the target region, and (4) the final gaze duration, or time interval between fixating inside the target region and making a response. In the most commonly observed search behavior, the eye enters the target region and the observer responds with a key press while looking at the target; segmentation of the elapsed time can be described, in terms of the above measures, as Overall RT = Initial Fixation + Scan time + Gaze Duration (Figure 2). Note that the measure of gaze duration inside the target region includes the motor response time. Analysis of variance statistics were performed on the mean reaction time for each subject, with a within-factor (epoch, from 1 to 10) and a between-factor (Consistent vs. Inconsistent groups). For the purpose of presenting the results, we restrict our discussion to the trials in which a target is present as this signal is the clearest indication of the role of identity priors in both experimental conditions. For simplicity, we treated two repetitions as one epoch for the graphs and statistical analyses detailed below.

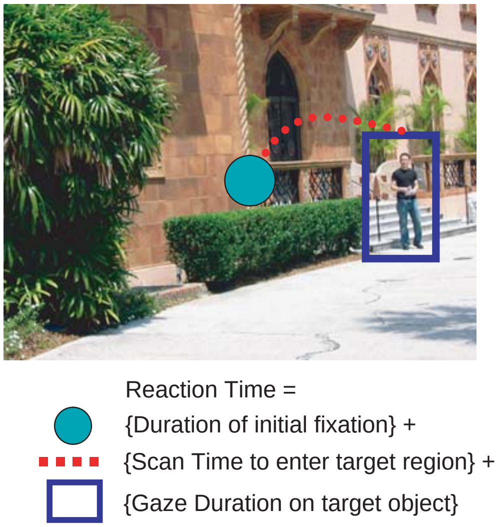

Reaction Time =
● {Duration of initial fixation} +
▪▪▪▪ {Scan Time to enter target region} +
□ {Gaze Duration on target object}

Figure 2: Illustration of the decomposition of the reaction time as the sum of the initial fixation, the scan time and the gaze duration on the target object.

2.4. Results

The data of two participants from the Consistent group were removed because of an eyetracker dysfunction. Of the remaining 30 observers, on average, participant's gaze entered the target region in 85% of consistent and 90% of inconsistent trials in which a target is present. These values do not change significantly across epoch (F<1). Statistical analyses of the overall RT, the initial fixation, the scan time, and the gaze duration are based on the target present trials in which the eye crossed into the target region, and for which overall RT remain below 3500 msec.

Mean Reaction Time: Figure 3 presents the mean RT data. As expected, learning occurs across repetitions in both conditions, but to a different final magnitude. The ANOVA (Epoch x Group) shows a significant effect of epoch (F(9,252)=35,p<.0001) and a significant interaction (F(9,252)=2.57, p<.01, i.e. the decrease in reaction time is steeper in the consistent group). When scenes are consistently associated with the presence or absence of a tar-

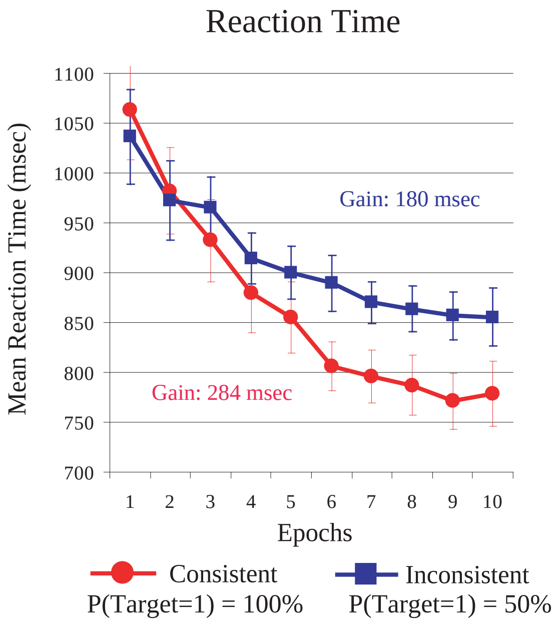


Figure 3: Mean Reaction Time for Consistent and Inconsistent condition, as a function of epochs.

get, learning occurs rapidly and efficiently (gain of about 280 msec by the 20th repetition). The improved search efficiency, however, is lower in the inconsistently mapped scenes (180 msec), where the target may or may not appear in a given scene background.

Initial Fixation Duration: The effect of learning on the duration of the initial scene fixation varies with scene prior strength. Figure 4A shows a clear initial fixation gain between the first and the 10th epoch in the inconsistent condition (F(9,15)=2.69, p<.01, this statistic is also validated with an item analysis, p<.0001). The duration of the initial fixation did not vary significantly with learning in the consistent condition (F~1).

Mean Scan Time: Scan time, the time from the end of the initial fixation to the entry in the target region, is plotted in Figure 4B. There is a very significant effect of learning on scan time duration (F9,252)=26.7, p<.0001), but no difference between the two groups (F<1): the delay before entering the target region decreases over repetitions equally in both the consistent and inconsistent conditions (~100 msec).

The invariance of the scenes themselves and the unchanging categorical priors suggest that building identity priors affects the search mechanism, and does so in an equal way for both inconsistent and consistent scenes. Note that given the increasingly efficient search, we might expect the total number of fixations to decrease over the course of the experiment. The decrease, though not dramatic, represents a statistically significant difference between the av-

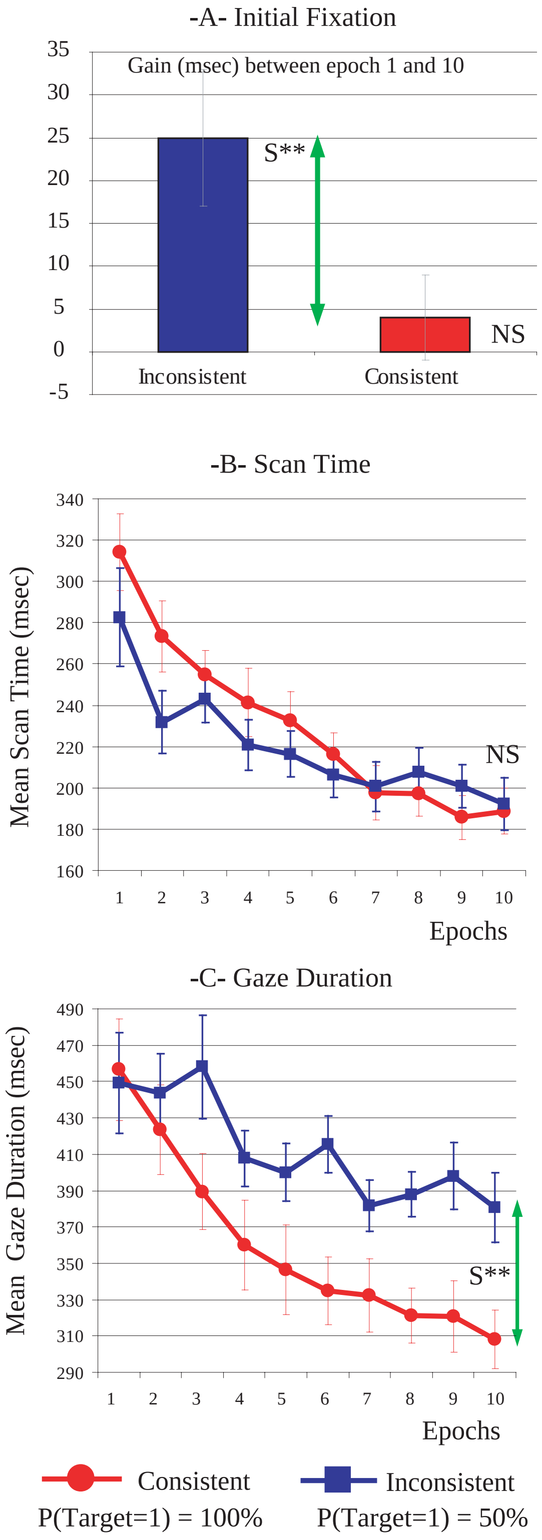


Figure 4: Decomposition of reaction time from Figure 3 for Consistent and Inconsistent conditions. -A- Gain in initial fixation, -B- Mean Scan Time, -C- Mean Gaze Duration. NS means not statistically significant. S** means the difference is statistically significant.

erage number of fixations per image for the initial and the final epochs in the consistent case (t(13)=5.8, p<.0001, from 3.3 to 2.8), but not in the inconsistent case (from an average of 3.2 to 3 fixations). The prominence of the target in the scene, as well as other local scene features (e.g. saliency regions, [6]), may play a role in holding the number of fixations almost constant for certain images. As such, number of fixations may not be the most precise measure of search efficiency, and analyses of fixations times as well as measurements of how direct vs. indirect the scan path is (cf. [1, 14]) will be the topic of a further work. In the present paper, overall scan time yields a more accurate measure of "search time" in that it reflects a continuous evaluation of overt target localization.

Mean Gaze Duration: The progression of gaze duration time is depicted in Figure 4C. Whereas scan time is related to the exploration phase of search, gaze duration is related to the recognition component of search (gaze duration is the duration of the fixation that occurs in the target region). This measure encompasses the period in which local image features of the region are evaluated and judged for compatibility with a target match, as well as motor responses. The ANOVA shows a main effect of groups (F(1,28)=4.8, p<.05) and epoch (F(9,252)=14.6, p<.0001) as well as a significant interaction (F(9,252)=2.06, p<.05). The gaze duration gain is about twice as much in the consistent case (148 msec) than the inconsistent case (68 msec).

2.5. Interpretation

How does specific scene context, in the form of well-rehearsed identity priors, exert its sphere of influence? Figure 5 summarizes how the total reaction time gain is distributed during the search process. In both conditions of learning, more than 98% of the overall gain in RT occurring between the first and the last (20th) repetition is explained by the decomposition of the search process into the three phases. The overall gain in the consistent condition (284 msec) is equally distributed between a gain in the scan time (exploratory search phase) and a gain in the gaze duration (target processing phase). The overall RT gain in the inconsistent condition (180 msec) is more sparsely distributed over the three stages of the search. A small but significant benefit (25 msec) in the duration of the initial fixation is observed over the course of scene context learning. In both conditions of learning, the cumulative gain in scan time is approximately equal (50% of the overall gain).

As expected, the identity priors of the inconsistent condition should serve chiefly as a cue of a target location, while accurate determination requires that the gaze approach the target region (93% of trials). In contrast, the scene recognition alone should be sufficient to diagnose target presence in the consistent condition. In theory no eye movements should be necessary to determine the presence of a target if the identity priors are fully used prior to programming the first saccade. However, we observe that the number of trials in which the subject's gaze enters the target region does not decrease significantly over time (average of 85%).

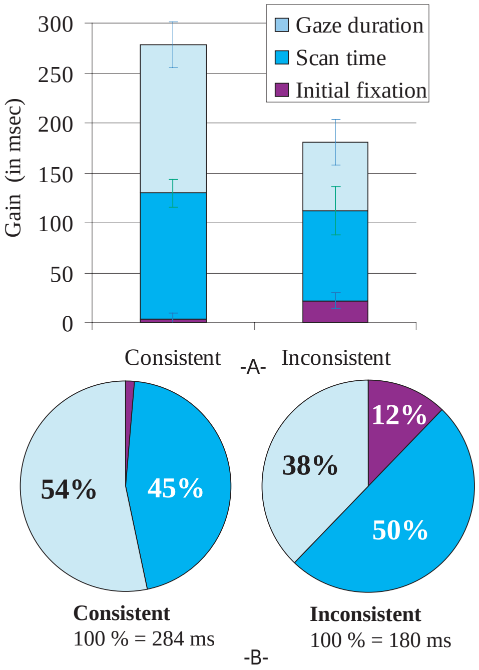


Figure 5: -A- Summary of the distribution of the overall reaction time gain (between the first and the 20th repetition). The overall RT gain is entirely explained as the sum of the gain occurring in the initial fixation, the scan time exploration and the gaze duration in the target region. -B- Breakdown of percentage of gain over the total learning period.

This suggests that identity prior information may not necessarily be used to guide search at an early stage. Other studies have shown that the memory-cued search may be more slowly activated than checking visual stimuli via covert or overt attention [9]. The initial saccade is still largely determined by the categorically-relevant region. However, the search exploration stage (scan time) and recognition stage (gaze duration on the target) are clearly influenced by the strength of the identity priors.

It is important to note that the interpretation of the gain in gaze duration cannot necessarily be attributed to the effect of identity priors on the ability to recognition the target. One possibility is that the strength of the identity priors reduces the sensory information required to process the target features. There is also the potential for automatic, or a priming effect, on motor planning to cause a similar, not easily separable, result. It is possible that the onset of a familiar scene may begin to initiate the planning of the correct response in the case where the scene itself is indicative of target presence or absence (consistent condition). This could also account for a greater gain in gaze duration in the consistent mapping condition. Additional experiments and scan path analyses will be needed to determine whether the gain observed in the target gaze duration should be attributed to a gain in the object processing time (perceptual level), or whether it reveals a benefit occurring at a decision stage (response priming in the consistent condition).

If identity priors are indeed directing attention as soon as the initial fixation, we may expect the effect to be observed, if not amplified, in the consistent condition. We observe, however, an absence of scene identity priors effects on the initial fixation latency in the consistent case. In this condition, the scene context itself carries information about the presence or absence of the target, in addition to the target location. This may suggest that when scene context is diagnostic of target presence, the primary role of the identity priors is to benefit the decision stage (responding about the presence of the object). This large benefit could be obscuring the early, comparatively small effect of covert attentional deployment (initial fixation duration). This is in fact what we observe in the consistent case: the reaction time benefit observed over twenty repetitions can be attributed to very large improvement in the two later stages of the search task. In addition to a significant decrease in scan time, the consistent scene-target mapping results in a decrease in the gaze duration on the target.

In the inconsistent condition, we observe that the three stages of the search process benefit from learning. The learning process is decreasing the delay of the initial fixation. In this case, the recognition of the scene does not predict the presence of the target (as it does in the consistent case). Because of target uncertainty, however, visually checking the target region is necessary on each trial to confidently determine whether the target is present. It is possible that observers have implicitly learned to systematically draw their attention towards the location of the target, and that it is this subtle scene guidance effect that is revealed in the shorter latency from the initial central region. The attentional draw of a particular location may have been emphasized in this condition because observers viewed each scene background twice as often as in the consistent case (see Figure 2).

3. Discussion

In this paper, we investigate the role of Identity priors in guiding visual search by monitoring eye movements as participants searched very familiar scenes for a target object. At which stages of visual search do contextual priors impact search performance?

One form of contextual guidance, termed a categorical prior, is likely to exert an early effect on a search task, in a pre-attentive manner, and before the first saccade is initiated. Evidence comes from findings from [10, 16] indicating that more than 80% of first eye fixations land within the categorically-specified contextual region. The effect of categorical prior [15] is illustrated in Figure 6: contextual scene information associated with the category of the scene modulates spatially-dependent expectations of target presence, thus driving the initial stage of the search mechanism.

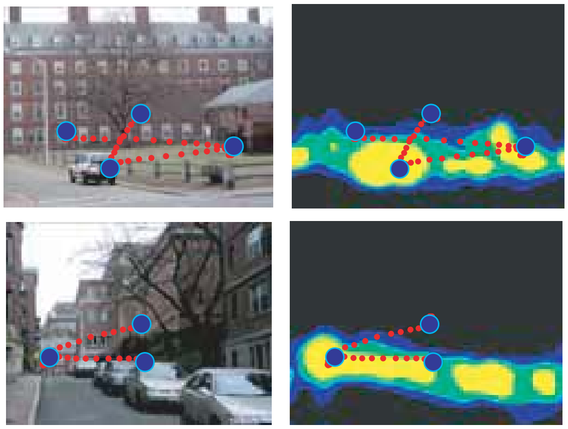
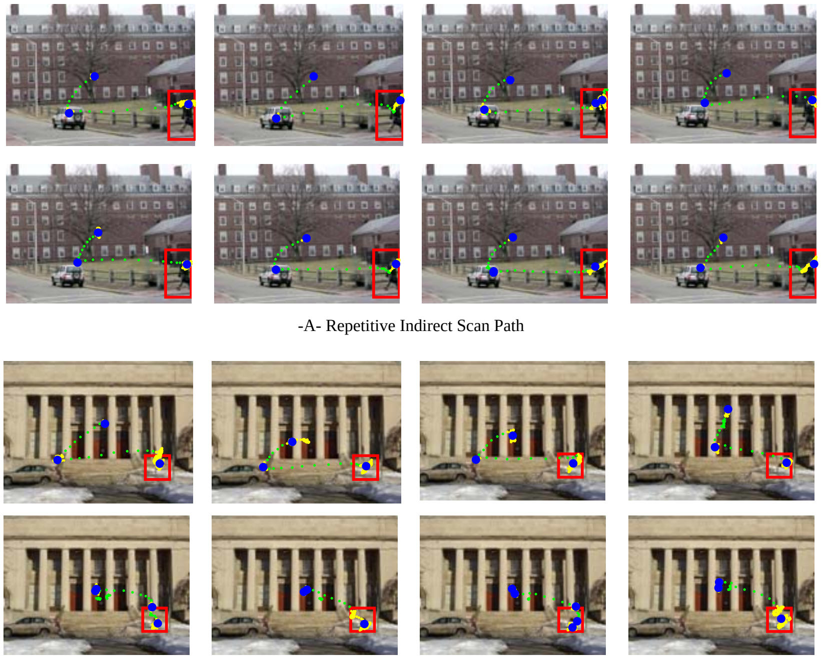


Figure 6: Illustration of the role of categorical prior on object search (based on the implementation of Torralba, 2003): fixations fall into the region defined by the categorical prior, here the association between urban scenes and probable locations of people. The fixation in the center corresponds to the initial fixation.

Our results show that the impact of identity priors on a search task is modulated by the strength of contextual relationship. After twenty repetitions of performing a search task, subjects establish a robust association between a particular scene and target, thereby improving their search performance. Although subjects improve their reaction time, eye movement patterns indicate that search is not very strongly influenced by immediate utilization of identity priors. Similarly, Chun and Jiang [2, 3] excluded the possibility that contextual cuing immediately guides attention to the scene-embedded target. Even though they found that responses to repeated configurations were faster than novel ones, the search slope for the repeated condition was never flat, as would be expected if the target was the first and only item examined. This is consistent with our observation of primary effects of identity priors on the scan time and gaze duration gain over scene context learning. Despite extensive learning, the fact that the number of fixations is not necessarily reduced to a single eye movement to the target region (see Figure 7) is potentially indicative of a non-zero search slopes, as in contextual cuing and repeated search task [2, 3, 9, 17]. This stands in contrast to the role of categorical priors, which guide overt attention to a contextually-relevant region at the earliest stages of search [10, 16].



-A- Repetitive Indirect Scan Path

-B- Indirect to Direct Scan Paths

Figure 7: Scan path examples illustrating two eye movement strategies. Fixations are depicted with blue dots. Eye movement trajectories are shown in green. Target region is represented by the red rectangle. -A- The effect of multiple scene exposures does not optimize the scan path from fixation to target. -B- Over several viewings of the scene, the trajectory from fixation to target region becomes increasingly optimized.

The nature of the role of context in search behavior remains largely unknown. In our experiment, scene recognition takes place as subjects examine the same stimuli over multiple repetitions and, in contrast to contextual cuing, the association is explicit. Peterson and Kramer [11] address the how contextual cuing influences eye movements and phrase the problem in terms of two components: recognition and guidance. They find that scene recognition could benefit search in a graded manner while guidance was always accurate. The features of the T and L stimuli used in their paradigm, however, differ significantly from the spatially diverse makeup of real-world scenes. The stimuli and paradigm used in our experiment provide an alternative model in which to study the nature of scene context on visual search, in particular an analytic decomposition of the time course of identity prior effects in ecologically-relevant search conditions.

Future studies will address the many open questions that remain in the aftermath of this initial study. Given the range of settings and types of target objects that comprise the real-world search environment, the effect of task difficulty on contextual guidance will be investigated. It is conceivable that contextual identity priors are activated to guide attention during a difficult search, while an easy search may rely on more image-based scene properties.

4. Conclusion

In the current study, investigations are described that begin to characterize the role of contextual identity priors in visual search of real-world scenes. A highly consistent relationship between a particular scene and target presence affords a greater overall search benefit than the condition in which identity priors are of intermediate strength. Analysis of eye movements localizes this benefit to result from a faster exploration stage (scan time) and a decrease in gaze duration on the target. Whereas the magnitude of target processing gain was dependent on the strength of the contextual relationship, the benefit to the exploratory search stage was independent of the strength of identity priors. As both categorical and identity context manifest themselves continuously in our interaction with the world, computational and behavioral approaches to human visual cognition should study the extent to which each type of contextual scene prior influences the allocation of attention in cognitive tasks, as well as the locus of the impact of such priors.

References

[1] Araujo, C., Kowler, E., & Pavel, M. Eye movements during visual search: the costs of choosing the optimal path. *Vision*

Research, Vol., 41, pp. 3613-3625, 2001.

[2] Chun, M. M., & Jiang, Y. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, Vol. 36, pp. 28-71, 1998.

[3] Chun, M.M., & Jiang, Y. Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, Vol. 10, pp. 361-365, 1999.

[4] Hayhoe, M., & Ballard, D. Eye movements in natural behavior. *TRENDS in Cognitive Sciences*, Vol. 9, pp. 188-194, 2005.

[5] Henderson, J.M., Weeks, P.A., & Hollingworth, A. Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 25, pp. 210, 1999.

[6] Itti, L., Koch, C., & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision*, Vol. 20(11), pp. 1254, 1998.

[7] Land, M.F., & Hayhoe, M.M. In what ways do eye movements contribute to everyday activities? *Vision Research*, Vol. 41, pp. 3559-3565, 2001.

[8] Logan, G.D. Towards an instance theory of automatization. *Psychological review*, Vol. 95, pp. 992-527, 1988.

[9] Oliva, A., Wolfe, J. M, & Arsenio, H. Panoramic Search: The interaction of Memory and Vision in Search through a Familiar Scene. *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 30, pp. 1132-1146, 2004.

[10] Oliva, A., Torralba, A., Castelhano, M. S. & Henderson, J. M. Top-Down control of visual attention in object detection. *Proc. IEEE Int. Conf. Image Processing*, Vol. 1, pp. 253-256, 2003.

[11] Peterson, M.S., & Kramer, A.F. Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception and Psychophysics*, Vol. 63, pp. 1239-1249, 2001.

[12] Schneider, W., & Shiffrin, R.M. Controlled and automatic human information processing. I. Detection, search and attention. *Psychological Review*, Vol. 84, pp. 1-66, 1977.

[13] Shinoda, H., HayHoe, M.M., & Shrivastava, A. What controls attention in natural environments? *Vision Research*, Vo. 41, pp. 3535-3545, 2001.

[14] Noton, D. & Stark, L. Scanpaths in Eye Movements during Pattern Perception. *Science*, New Series, Vol. 171, pp. 308-311, 1971.

[15] Torralba, A. Modeling global scene factors in attention. *J. Opt. Soc. Am. A: Special Issue on Bayesian and Statistical Approaches to Vision*, Vol. 20, pp.1407-1418, 2003.

[16] Torralba, A., Oliva, A., Castelhano, M.S., & Henderson, J.M. Saliency, objects and scenes: global scene factors in attention and object detection. *Journal of Vision*, Vol. 4(8), pp. 337a, 2004.

[17] Wolfe, J.M., Klempen, N., & Dahlen, K. Post-attentive Vision. *The Journal of Experimental Psychology: Human Perception and Performance*, Vol. 26, pp. 693-716, 2000.