The Attraction of Visual Attention to Texts in Real-World Scenes

Hsueh-Cheng Wang (hchengwang@gmail.com)
Marc Pomplun (marc@cs.umb.edu)
Department of Computer Science, University of Massachusetts at Boston,
100 Morrissey Boulevard, Boston, MA 02125 USA

Abstract

Intuitively, it seems plausible that in real-world scenes, attention is disproportionately attracted by texts. The present study tested this hypothesis and examined some of the underlying factors. Texts in real-world scenes were compared with paired control regions of similar size, eccentricity, and low-level visual saliency. The greater fixation probability and shorter minimum fixation distance of texts showed their higher attractiveness. These results might be caused by the prominent locations or special visual features of text. In another experiment, texts were removed from the scenes, and the results indicated that the locations that used to contain texts did draw more attention than controls. Finally, texts were placed in unexpected positions in front of homogeneous and inhomogeneous backgrounds. These unconstrained texts were found more attractive than controls, with background noise reducing this difference, which indicates that the attraction by specific visual features of text was superior to typical saliency.

Keywords: real-world scenes; scene syntax; text; eye movements; visual attention; LabelMe.

Introduction

When inspecting real-world scenes, human observers continually shift their gaze to retrieve information. Important pieces of information could be, for instance, depictions of objects (e.g., cars, monitors, or printers) or texts, which could be shown on depictions of signs, banners, license plates, and other objects. Observers' attention has been found to be biased toward visually salient locations, e.g., high-contrast areas, during scene viewing or search (Itti & Koch, 2001). Torralba, Oliva, Castelhano, and Henderson (2006) suggested that scene context, i.e., the combination of objects that have been associated over time and are capable of priming each other to facilitate object and scene categorization, predicts the image regions likely to be fixated. Võ and Henderson (2009) claimed that scene syntax, i.e., the position of objects within the specific structure of scene elements, influences eye movement behavior during real-world scene viewing.

It is still an open question whether texts in real-world scenes attract more attention than comparable regions and why this would be the case. It is possible that low-level visual saliency attracts attention, i.e., that texts are more attractive because they typically carry higher saliency – as computed along the lines of Itti and Koch (2001) - or luminance contrast. Moreover, it is also possible that the positions of texts are more predictable in the scene context to contain important information, for example, texts on street signs. Such an effect would be in line with the model by Torralba et al. (2006), which predicts the image regions likely to be fixated in a natural search task based on the expected location of the target.

The goal of the present study was to investigate the contribution of low-level visual saliency and high-level features to the ability of texts to attract attention in real-world scene viewing. To test if texts are more attractive than other scene objects, an eye-tracking database of scene viewing by Judd, Ehinger, Durand, and Torralba (2009) was re-analyzed in Experiment 1.

Experiment 1: Reanalysis of Previous Data

Method

Participants. Judd and colleagues (2009) collected eye tracking data of 15 viewers. These viewers were males and females between the ages of 18 and 35. Two of the viewers were researchers on their project and the others were naive viewers.

Apparatus. All viewers sat at a distance of approximately two feet from a 19-inch computer screen of resolution 1280×1024 in a dark room and used a chin rest to stabilize their head. An eye tracker with the sampling rate of 240 Hz recorded their eye movements on a separate computer**.**

Procedure. All participants freely viewed each image for 3 seconds, separated by 1 second of viewing a gray screen. To ensure high-quality tracking results, camera calibration was checked every 50 images. All images were divided into two sessions of 500 randomly ordered images. Each session was done on average at one week apart. After every 100 images being presented, participants were asked to indicate which images they had seen before to motivate them to pay attention to the images

Stimuli. There were 1003 images in the database by Judd et al. (2009), and these images included both outdoor and indoor scenes. Some of these images were selected from the LabelMe database (see below).

Analysis. To identify and localize text in real-world scene stimuli, we used the freely available LabelMe image dataset (Russell, Torralba, Murphy & Freeman, 2008) containing a large number of scene images that were manually segmented into annotated objects. The locations of objects are provided as coordinates of polygon corners and are labeled by English words or phrases. Out of the 1003 images we selected 57 images containing 240 text-related labels and 93 images containing non-text objects. The text-related labels included terms such as 'text', 'banner', or ‘license plate’. For the non-text objects, we excluded objects with text-related labels or background labels, e.g., ‘floor’, ‘ceiling’, ‘wall’, ‘sky’, ‘crosswalk’, ‘ground’, ‘road’, ‘sea’, ‘sidewalk’, ‘building’, or ‘tree’. The label 'face' was also excluded since faces have been shown to be particularly attractive (see Judd et al., 2009, for a review). There were 1620 non-text objects in the final selection. The resolution of these images was adjusted to 1024×768 pixels, and the coordinates of all objects were updated accordingly.

The raw eye movement data was smoothed using a computer program developed by Judd et al. (2009) that calculates the running average over the last 8 data points (i.e., over a 33.3 ms window). A velocity threshold of 6 degrees per second was used for saccade detection. Fixations shorter than 50 ms were discarded (see Judd et al., 2009).

It is known that *eccentricity* (the distance between the center of an object to the center of the screen) and *size* (number of pixels) of an object might influence eye movement measures. Observers show a tendency to fixate near the center of the screen when viewing scenes on computer monitors (Tatler, 2007). Larger objects tend to be fixated more frequently since the landing probability increases with larger area. Low-level visual features such as *saliency* and *luminance contrast* were computed. Saliency was calculated by the freely available computer software “Saliency Map Algorithm” using the standard Itti, Koch, and Niebur (1998) saliency map based on color, intensity, orientation, and contrast. The average saliency value of pixels inside an object boundary was used to represent object saliency. Luminance contrast was defined as the gray-level standard deviation of pixels enclosed in an object. On average, text objects occupied 1.43% of the area in a 1024×768 pixel display.

To derive compatible control objects, non-text objects were binned by eccentricity (smaller than 200, between 200 and 300, and greater than 300) and size (smaller than 1650, between 1650 and 5600, and greater than 5600). These ranges of eccentricity and size were selected to roughly include the same number of objects in each interval. Each text object was paired with one non-text object within the same size and eccentricity interval and matched in terms of saliency and luminance contrast as closely as possible. A text object and its non-text match were typically selected from different images.

Additionally, for each text object a control region in the same scene was set up that matched its counterpart exactly in its shape and size and had similar eccentricity (Ecc.), saliency (Sal.), and luminance contrast (LumC.) (see Figure 1). The control regions could enclose non-text objects or backgrounds but did not intersect with any text objects. The characteristics of text objects, non-text objects, and control regions (Con. Region) are summarized in Table 1.

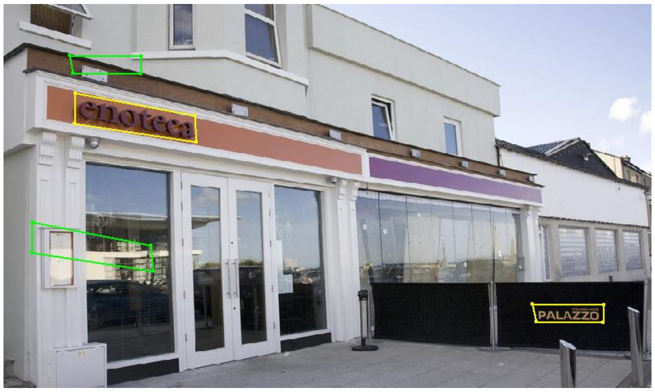

Figure 1. Texts (yellow polygons) and their paired control regions (green polygons) in one of the scene stimuli.

Table 1: Average characteristics of text objects, non-text objects, and control regions.

	Size	Ecc.	Sal.	LumC.
Experiment 1				
Text	2631	283	0.41	40
Non-Text	2828	292	0.41	40
Con. Region	2631	283	0.37	46
Experiment 2				
Erased Text	2631	283	0.43	21
Non-Text	2676	293	0.41	24
Con. Region	2631	283	0.37	36
Experiment 3				
UncText H B	2351	288	0.20	10
UncText INH B	2723	281	0.39	55
UncText H	2351	288	0.24	34
UncText INH	2723	281	0.40	57
Non-Text H	2670	301	0.27	34
Non-Text INH	2746	284	0.41	57
Con. Region H	2351	287	0.28	40
Con. Region INH	2723	281	0.41	55

In order to measure the attraction of visual attention, two object-based eye movement measures were taken: *fixation probability* (the probability of a fixation to land inside a text or non-text object or a control region during a trial) and *minimum fixation distance* (the shortest Euclidean distance from the center of the object to any fixation during a trial). If an object had higher fixation probability or shorter minimum fixation distance, the object was considered more attractive. If there was no fixation landing inside an object boundary, its fixation probability was 0 regardless of how close a fixation approached the object. Minimum fixation distance was measured to overcome this drawback and provide convergent evidence for any attractiveness results.

Results and Discussion

Fixation probability and minimum fixation distance of texts, non-texts and control regions are shown in Figure 2. The fixation probability of texts was significantly higher than the one of non-text objects and control regions, both Fs(1; 14) > 76.85, ps < 0.001. Minimum fixation distance of texts was shorter than the one of non-text objects and control regions, both Fs(1; 14) > 46.53, ps < 0.001. Both results were consistent and suggested that texts were more attractive than both non-text objects and control regions. In addition, non-text objects had higher fixation probability and shorter minimum fixation distance than control regions, both Fs(1; 14) > 45.15, ps < 0.001. The results might be caused by control regions not having an obvious boundary like texts and non-text objects.

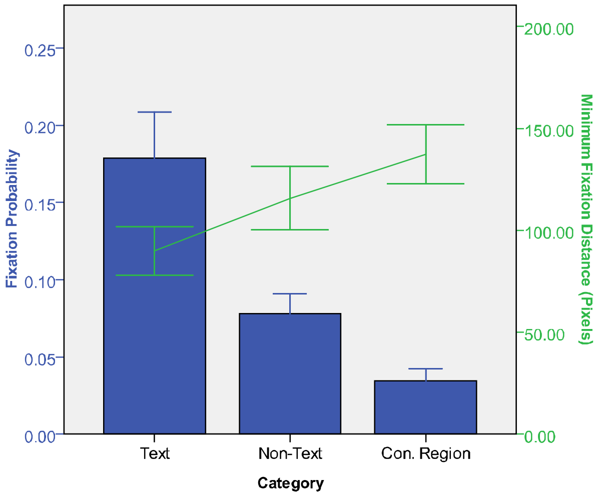

Figure 2. Fixation probability and minimum fixation distance of texts, non-texts, and control regions. In this chart and all following ones, error bars are based on 95% confidence intervals.

The observed effect might be caused by low-level visual saliency as computed by the Saliency Map Model (Itti & Koch, 1998), high-level features (expected locations), or maybe unique visual features of texts. Texts, like faces, might have their unique visual features that are unrelated to typical low-level visual saliency so that human observers develop "text detectors" during everyday scene viewing. The selected controls ruled out the first hypothesis of low-level visual saliency. We will test how expected locations affect eye movements in Experiment 2, and the influence of unique visual features of texts on attention will be examined in Experiment 3.

Experiment 2: Erased Text

To test whether the locations of text placement contribute to the attractiveness of texts, in Experiment 2 we “erased” the text parts from text objects and examined whether the observers’ attention was still biased toward these objects. The text removal sometimes causes strong oddness, e.g., for a stop sign, but sometimes does not, such as for a billboard. This oddness is due to viewers expecting text in that location, which might possibly attract more attention.

Method

Participants. Fifteen participants performed this experiment. All were students at the University of Massachusetts Boston, aged between 19 to 40 years old, and had normal or corrected-to-normal vision. Each participant received 10 dollars for participation in a half-hour session.

Apparatus. Eye movements were recorded using an SR Research EyeLink-II system with a sampling frequency of 500 Hz. After calibration, the average error of visual angle in this system is 0.5˚. Stimuli were presented on a 19-inch Dell P992 monitor with a refresh rate of 85 Hz and a screen resolution of 1024×768 pixels. Participants’ responses were entered using a game-pad**.**

Procedure. After participants read the instructions, a standard 9-point grid calibration (and validation) was completed. Following two practice trials, participants viewed 130 stimuli in random order. They were instructed to freely inspect the scene. At the start of each trial, a drift calibration screen appeared, and participants were instructed to look at the calibration dot that appeared in the center of the screen. After subjects had passed the drift correction, the stimuli were presented. Following a ten-second presentation of each scene, the stimulus disappeared and the calibration dot appeared again. In some cases, calibration and validation were performed once again to increase eye-tracking accuracy**.**

Stimuli. The same 57 images and 240 text regions used in Experiment 1 were employed in Experiment 2. However, in Experiment 2, the “text parts” in text objects were removed manually by replacing them with the background color of the texts as shown in Figure 3. This removal led to a reduction in average luminance contrast from 40 to 21 (see Table 1). Nonetheless, the average saliency was not affected by this text removal, due to the computation of saliency being based on center-surround differences in color, intensity, and orientation. Note that luminance contrast was computed exclusively within an object, but saliency was calculated according to the whole image, and the neighbor regions of an object were taken into account. Therefore, a stop sign might still be salient without the text “stop” because of the color difference between the sign and its surroundings but its luminance contrast is reduced since there is no contrast inside the sign.

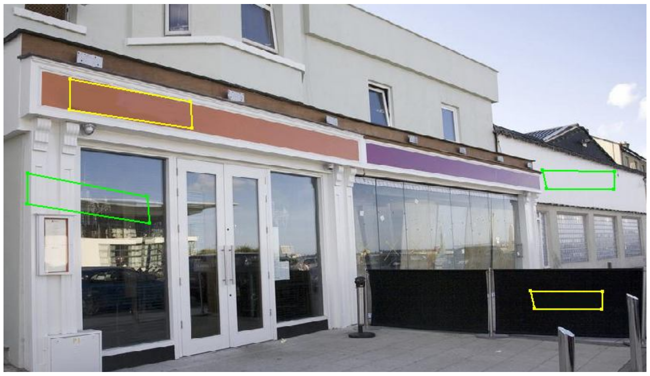

Figure 3. Erased texts and their paired control regions in a scene.

Analysis. The raw eye movement data were parsed using the standard EyeLink algorithm. Eye fixation data were analyzed separately for the first 3 seconds and for the entire 10-second viewing duration. Since this study did not involve any post-presentation questions, the first 3 seconds of viewing should be comparable with the total 3 seconds of viewing in Experiment 1. As described in Experiment 1, non-text objects and control regions were chosen based on similar size, eccentricity, saliency, and luminance contrast (see Table 1). Since saliency and luminance contrast were positively correlated, r = 0.34, luminance contrast of control regions (36) was higher than that of removed-text regions (21).

Results and Discussion

As shown in Figure 4, for 3-second viewing in Experiment 2, fixation probability for erased texts dropped compared to text objects in Experiment 1, F(1; 28) = 35.82, p < 0.001, for between-subject ANOVA. Minimum fixation distance for erased texts was significantly longer in Experiment 2 than for texts in Experiment 1, F(1; 28) = 10.53, p < 0.01 (see Figure 5). These results might be caused by the reduction of saliency and luminance contrast that accompanied the erasure of text.

During 3- and 10-second viewing, erased texts had slightly higher fixation probability than non-text objects, but this difference was not statistically significant, all Fs < 1, ps > 0.33. However, minimum fixation distance for missing texts was shorter than for non-text objects during 3-second viewing, F(1; 14) = 25.57, p < 0.001, and 10-second viewing, F(1; 14) = 14.43, p < 0.01, showing that the typical locations of text still matter even when they do not contain any text. This result indicates that part of the attractiveness of texts derives from their prominent, expected locations in typical real-world images. To test how the unique visual features of texts attract attention without the effects of expected locations, Experiment 3 dissociated texts from their typical locations and placed them in front of homogeneous or inhomogeneous background. The purpose of using inhomogeneous backgrounds was to reduce the unique visual features of text by adding visual noise, and we expected to find less attraction of attention by texts in front of such inhomogeneous background.

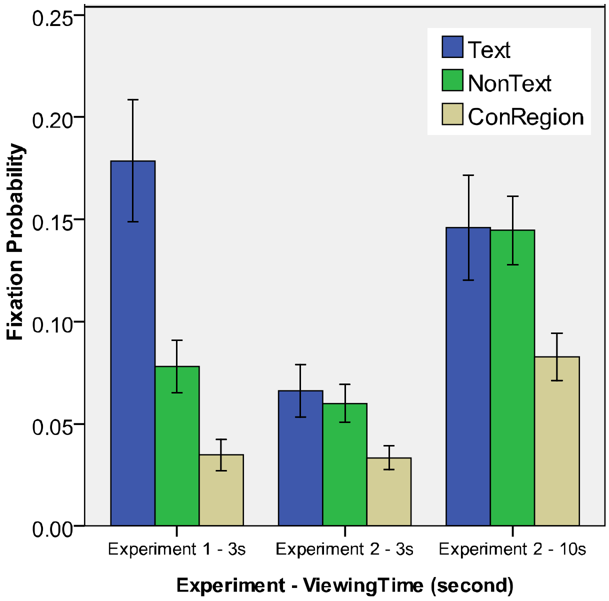

Figure 4. Fixation probability of texts in Experiment 1, erased texts in Experiment 2, and non-texts and control regions in both experiments.

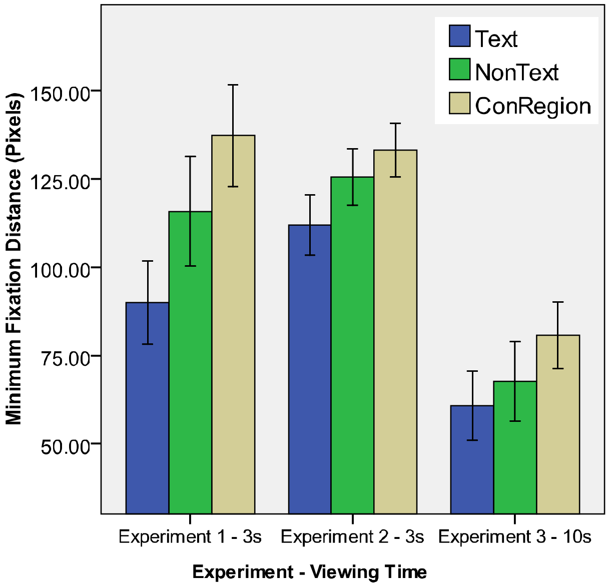

Figure 5. Minimum fixation distance of texts in Experiment 1, erased texts in Experiment 2, and non-texts and control regions in both experiments.

Experiment 3: Unconstrained Text

Experiment 3 "moved" the text parts to unexpected locations and placed them on high or low luminance contrast backgrounds. This design eliminated the influence of expected locations and tested how the unique visual features of text affected eye movements.

Method

Participants. An additional 15 students from the University of Massachusetts at Boston participated in this experiment. None of them had participated in Experiment 2.

Apparatus. Eye movements were recorded using an SR Research EyeLink Remote system. Other settings were the same as in Experiment 2**.**

Procedure. The procedure was identical to Experiment 2**.**

Stimuli. To extract the "text part" of a text object, the difference in each of the RGB color components of every pixel in each text object between Experiments 1 and 2 were calculated. These patterns of color differences were recreated in other, randomly chosen scenes and placed in positions where the original size and eccentricity were maintained (see Figure 6). These unconstrained texts were prevented from overlapping with regions currently or previously occupied by texts. There were a total of 240 unconstrained text objects. Half of them were placed on homogeneous background, i.e., regions with the lowest luminance contrast of all possible locations before placing the text parts, while the others were placed on inhomogeneous background, i.e., those areas with the highest luminance contrast. To prevent an unconstrained text from being placed on a computationally inhomogeneous but visually homogeneous background, e.g., half black and half white, the luminance contrast of a candidate region was calculated using 10×10 pixels windows covering the candidate region.

As discussed above, inhomogeneous backgrounds might cause visual noise that interferes with the unique visual features of texts and thereby reduces the attraction of the viewers' attention by such features. Table 1 shows the characteristics of the unconstrained text in front of homogeneous background before (UncText H B) and after (UncText H) the text parts were placed as well as the unconstrained texts in front of inhomogeneous background before (UncText INH B) and after (UncText INH) the text parts were placed.

Analysis. The analyses were identical to Experiment 2. Both 3- and 10-second viewing durations were analyzed for unconstrained texts in front of homogeneous and inhomogeneous backgrounds. Each unconstrained text was paired with a non-text object and a control region using the same methods applied in Experiments 1 and 2. Table 1 lists the characteristics of paired non-text objects and control regions.

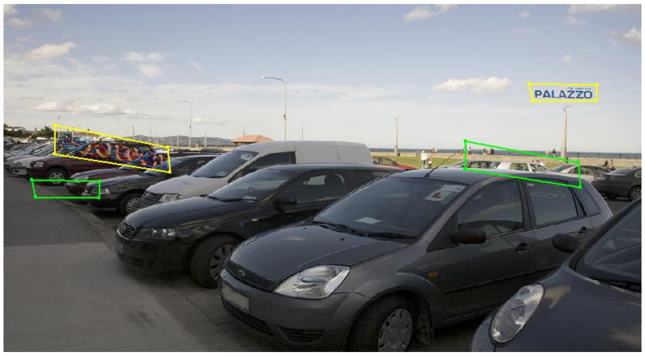

Figure 6. Unconstrained texts (yellow polygons) in front of homogeneous (right) and inhomogeneous backgrounds (left) and their paired control regions (green polygons) in one of the scene stimuli.

Results and Discussion

As shown in Figure 7, the fixation probability of unconstrained texts in front of homogeneous background was higher than for non-texts and control regions during 3-second viewing, both Fs(1; 14) > 34.98, ps < 0.001. The unconstrained texts in front of homogeneous background (mean fixation probability: 0.18) were as attractive as texts in Experiment 1 (mean fixation probability: 0.18) located in expected positions, F = 0.01, p > 0.9. For unconstrained texts in front of inhomogeneous background, the fixation probability was still significantly higher than for non-texts and control regions, both Fs(1; 14) > 14.76, ps < 0.01, but the difference was not as large as for unconstrained texts in front of homogeneous background. Although saliency (0.40) and luminance contrast (57) of inhomogeneous background were higher than the ones of homogeneous background (0.24 and 34, respectively), this result suggests that inhomogeneous background caused noise that interfered with identifying the distinctive visual features of texts. For 10-second viewing, the fixation probability for unconstrained texts in front of both homogeneous and inhomogeneous background as well as for their control regions increased. These results were identical to 3-second viewing.

For minimum fixation distance, the trends were similar to fixation probability; unconstrained texts in front of homogeneous and inhomogeneous background received shorter distances and can therefore be considered more attractive (see Figure 8). Minimum fixation distance of unconstrained texts in front of homogeneous background was significantly higher than that of their associated non-text objects and control regions, both Fs(1; 14) > 7.66, ps < 0.05. However, the corresponding comparisons for inhomogeneous background failed to reach significance. For 10-second viewing, minimum fixation distances of all categories were reduced and the results were similar to what was found during 3-second viewing.

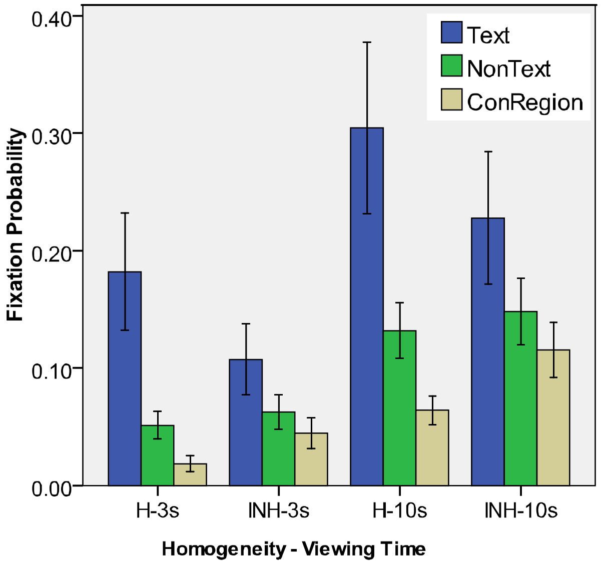

Figure 7. Fixation probability of unconstrained texts in front of homogeneous (H) and inhomogeneous (INH) background, and the corresponding non-text objects and control regions.

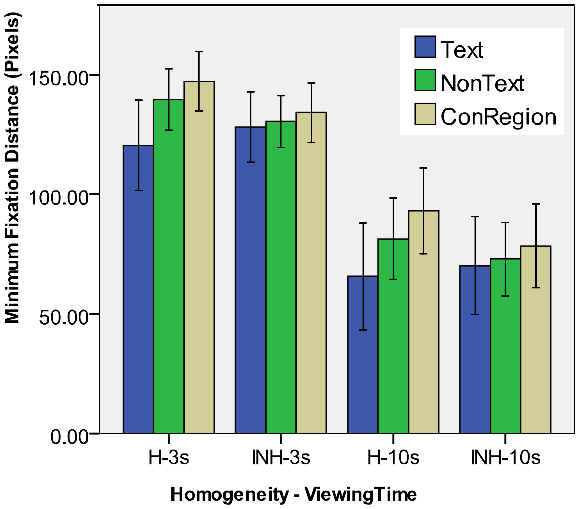

Figure 8. Minimum fixation distance of unconstrained texts in front of homogeneous (H) and inhomogeneous (INH) background, non-text objects, and control regions.

General Discussion

In Experiment 1, we found that text objects were more attractive than non-text objects and control regions of similar size, eccentricity, saliency, and luminance contrast. Since we controlled for the typical saliency computed by color, intensity, orientation, and contrast, the results might be caused by high-level features (expected locations), special visual features of text, or both. Experiment 2 further investigated the attraction of attention by high-level features, and the results suggested that eye fixations were influenced by expected locations that might possibly be more informative. This finding has important implications for our understanding of attention in real-world scenes. First, it supports the concept of "contextual guidance" found by Torralba et al. (2006). Second, and most importantly, it demonstrates that this factor does not only apply to search tasks but that expected locations play a role even in a free viewing task. By presenting the unique visual features of text in unexpected locations and in both fully visible and degraded variants, the results of Experiment 3 indicated that the specific visual features of texts were superior to typical saliency, and their influence on attention was reduced by the noise caused by inhomogeneous background. We conclude that both low- and high-level features contribute to the ability of texts to attract a disproportionate amount of visual attention in real-world scenes. However, the results obtained in Experiment 3 might also be caused by the replacement of texts inducing semantic or syntactic violation. To further investigate the special visual features of texts, the next step in this line of research could be an experiment that places non-text objects in unexpected locations. In addition, it is important to investigate the contribution of informativeness to the ability of texts to attract attention. Such experiments could present, for instance, non-English texts, such as Chinese characters, to native English speakers as subjects.

Acknowledgments

Parts of the data were presented at the European Conference on Visual Perception (ECVP 2010). Thanks to Melissa Võ for helpful comments on an earlier version of the article. Preparation of the article was supported by Grant R15EY017988 from the National Eye Institute to Marc Pomplun.

References

Itti, L, Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans Pattern Analysis and Machine Intelligence* 20 (11): 1254-1259.

Itti, L., & Koch, C. (2001). Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*. 2(3):194-203.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look, *IEEE International Conference on Computer Vision (ICCV).*

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation, *International journal of computer vision*, 77, 1-3, 157-173.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 1-17.

Torralba, A., Oliva, A., Catelhano, M., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113, 766-786.*

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3):24, 1-15.