The attraction of visual attention to texts in real-world scenes

Hsueh-Cheng Wang

Department of Computer Science, University of Massachusetts at Boston, Boston, MA, USA

Marc Pomplun

Department of Computer Science, University of Massachusetts at Boston, Boston, MA, USA

When we look at real-world scenes, attention seems disproportionately attracted by texts that are embedded in these scenes, for instance, on signs or billboards. The present study was aimed at verifying the existence of this bias and investigating its underlying factors. For this purpose, data from a previous experiment were reanalyzed and four new experiments measuring eye movements during the viewing of real-world scenes were conducted. By pairing text objects with matching control objects and regions, the following main results were obtained: (a) Greater fixation probability and shorter minimum fixation distance of texts confirmed the higher attractiveness of texts; (b) the locations where texts are typically placed contribute partially to this effect; (c) specific visual features of texts, rather than typically salient features (e.g., color, orientation, and contrast), are the main attractors of attention; (d) the meaningfulness of texts does not add to their attentional capture; and (e) the attraction of attention depends to some extent on the observer's familiarity with the writing system and language of a given text.

Keywords: real-world scene, scene viewing, text attraction

Citation: Wang, H.-C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision, 12*(6):26, 1–17, http://www.journalofvision.org/content/12/6/26, doi:10.1167/12.6.26.

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, printers) or texts, which could be shown on depictions of signs, banners, advertisement billboards, license plates, and other objects. Human text detection in natural scenes is critically important for people to survive in everyday modern life, for example, by drawing attention to traffic signs or displays showing directions to a hospital or grocery store.

Texts in real-world scenes were found to attract more attention than regions with similar size and position in a free viewing task (Cerf, Frady, & Koch, 2009), but it is still an open question as to what factors would control such an attentional bias toward texts. It is possible that low-level visual *saliency* attracts attention (e.g., Itti, Koch, & Niebur, 1998; Bruce & Tsotsos, 2006; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002), or top-down control of visual attention (e.g., Hwang, Higgins, & Pomplun, 2009; Peters & Itti, 2007; Pomplun, 2006; Zelinsky, 2008). Another possibility is that texts typically carry higher saliency, luminance contrast, or edge information. Baddeley and Tatler (2006) suggested that different attention-attracting features are likely correlated in natural scenes (e.g., high spatial frequency edge-content information is typically associated with high contrast), and they found that edge-content information predicts the positions of fixations more accurately than do other features, such as contrast. The edge measures may thus be important factors that make texts more attractive than other objects.

When studying the allocation of visual attention, it is important to consider the relative contributions of objects and low-level features. Elazary and Itti (2008) used the LabelMe image dataset (Russell, Torralba, Murphy, & Freeman, 2008) to examine the relation between objects and low-level saliency, as computed by the model of Itti et al., and they found that salient locations tend to fall within "interesting objects" defined by objects people choose to label. Their finding was later refined by Nuthmann and Henderson (2010), who showed that viewers tend to fixate close to the center of objects and emphasized the importance of objects in memorization and preference tasks. Einhäuser, Spain, and Perona (2008) further investigated whether observers attend to interesting objects by asking them to name objects they saw in artistic evaluation, analysis of content, and search tasks. Einhäuser et al. (2008) found that saliency combined

doi: 10.1167/12.6.26 Received November 22, 2011; published June 19, 2012 ISSN 1534-7362

with object positions determines which objects are named frequently. They concluded that both low-level saliency and objects need to be integrated in order to capture attention.

Furthermore, attentional capture could be driven by some particular classes of objects, which attract eye fixations independently of their low-level visual saliency. There may be specific features of texts, similar to faces that attract attention but differ from those features that are typically associated with visual saliency. For instance, Cerf, Cleary, Peters, Einhäuser, and Koch (2007) showed that a model combining low-level saliency and face detection achieved better estimation of fixation locations than low-level saliency alone. Similarly, Judd, Ehinger, Durand, and Torralba (2009) added object detectors for faces (Viola & Jones, 2004) and persons (Felzenszwalb, McAllester, & Ramanan, 2008) to their model and obtained better prediction of human fixations. Cerf et al. (2009) refined the standard saliency model by adding a channel indicating regions of faces, texts, and cell phones, and demonstrated that the enhancement of the model significantly improved its ability to predict eye fixations in natural images.

Moreover, it is also possible that the typical *locations* of texts in the scene context are more predictable to contain important information and thus attract a disproportionate amount of attention. 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. Furthermore, 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. Such an effect would be in line with the studies of dependency among objects (e.g., the relative position of a plate and silverware; Oliva & Torralba, 2007) and the contextual guidance model (Torralba et al., 2006), which predicts the expected location of the target in a natural search task based on global statistics from the entire image. Furthermore, Eckstein, Drescher, and Shimozaki (2006) recorded viewers' first saccades during a search for objects that appeared in expected and unexpected locations in real-world scenes, and they found the endpoints of first saccades in target-absent images to be significantly closer to the expected than the unexpected locations. Adding to the above results, an experiment by Mack and Eckstein (2011) investigated the effects of object co-occurrence on visual search, and it was found that viewers searched for targets at expected locations more efficiently than for targets at unexpected locations.

Finally, the *familiarity* of texts to viewers might also influence the attractiveness of texts; for example, observers' attention may or may not be attracted by the contents of an information board in a language that they do not understand. Cerf et al. (2009) implied that attention to text may be developed through learning. If this assumption holds, a writing system familiar to viewers would be expected to catch more attention than an unfamiliar one. Here we have to distinguish between the meaning of texts that is inaccessible to observers who do not speak the given language, and their potential unfamiliarity with the writing system, i.e., the visual features of the texts. Both factors need to be investigated separately.

The goal of the present study was to investigate the contributions of low-level visual saliency, expected locations, specific visual features, and familiarity of texts to their ability to attract attention in real-world scene viewing. In order to test if texts are more attractive than other scene objects, in Experiment 1 an eye-tracking database of scene viewing by Judd et al. (2009) was first reanalyzed. In Experiments 2 to 5, new eye-movement data were collected and analyzed to study the factors that underlie the attraction of attention by texts.

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 2 feet from a 19-inch computer screen of resolution 1,280 × 1,024 in a dark room and used a chin rest to stabilize their head. A table-mounted, video-based ETL 400 eye tracker (ISCAN Inc., Woburn, MA) with a sampling rate of 240 Hz recorded their eye movements using a separate computer (Judd et al., 2009). The images were presented at approximately 30 pixels per degree.

Stimuli

There were 1,003 images in the database by Judd et al. (2009), and these images included both outdoor and indoor scenes. Some of these images were included in the freely available LabelMe image dataset (Russell et al., 2008) which contains 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.

Procedure

All participants freely viewed each image for 3 seconds, separated by 1 second of viewing a gray, blank 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. The two sessions were done on average at one week apart. After the presentation of every 100 images, participants were asked to indicate which images they had seen before to motivate them to pay attention to the images.

Analysis

The LabelMe dataset was used to identify and localize text in real-world scene stimuli. Out of the 1,003 images, we selected 57 images containing 240 text-related labels and another 93 images containing only non-text objects. Figure 1a shows one of the scene stimuli containing texts. 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' since previous research has shown that viewers prefer looking at objects over background (Buswell, 1935; Henderson, 2003; Yarbus, 1967; Nuthmann & Henderson, 2010). It must be noted that the definition of background is not entirely clear (Henderson & Ferreira, 2004). For example, objects labeled as 'building' or 'tree' may or may not be considered as background. To reduce the ambiguity, this study excluded 'building' and 'tree' from the set of non-text objects. 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 1,620 non-text objects in the final selection. The images were rescaled to have a resolution of 1,024 × 768 pixels (approximately 34×26 degrees of visual angle), and the coordinates of all objects were updated accordingly.

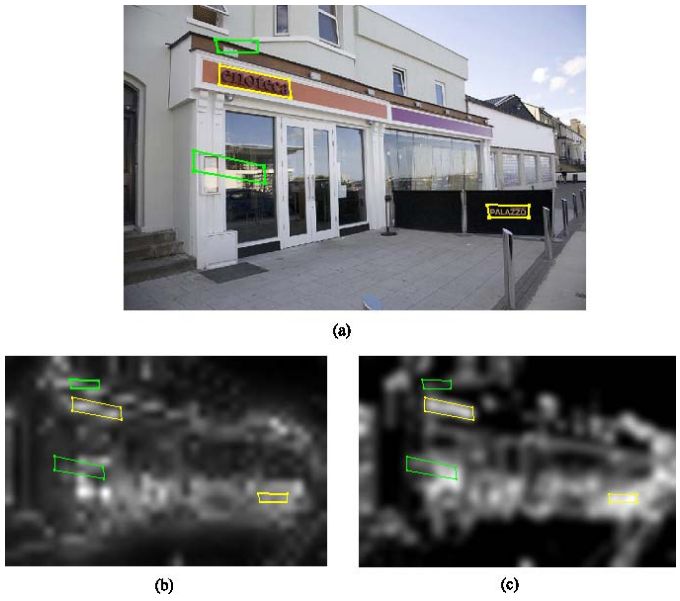


Figure 1. (a) Texts (yellow polygons) and their paired control regions (green polygons) in one of the scene stimuli. The corresponding saliency and edge-content information are illustrated in (b) and (c).

The raw eye movement data were 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 (Judd et al., 2009).

In the analysis, several variables needed to be controlled for, such as the *eccentricity* and *size* of objects. It is known that these variables influence eye-movement measures, because observers tend to fixate near the center of the screen when viewing scenes on computer monitors (Tatler, 2007) and larger objects tend to be fixated more frequently. The eccentricity of an object (the distance from its center to the center of the screen) and its size (number of pixels) were calculated according to the coordinates provided by LabelMe. In order to control for low-level visual features in our analyses, we computed saliency, luminance contrast, and edge-content information of LabelMe objects. Saliency was calculated by the freely available computer software "Saliency Map Algorithm" (http://www.klab.caltech.edu/~harel/share/gbvs.php, retrieved on December 25, 2011) by Harel, Koch, and Perona (2006) using the standard Itti, Koch, and Niebur (1998) saliency map based on color, intensity, orientation, and contrast as shown in Figure 1b. 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. For computing edge-content information, images were convolved with four Gabor filters, orientated at 0, 45, 90, and 135 degrees. Tatler et al. (2005) suggested to set the spatial frequency of the Gabor carrier to values between 0.42 and 10.8 cycles per degree, and we chose a value of 6.75 cycles per degree. All computations followed Tatler et al. (2005) and Baddeley and Tatler (2006) except that a popular boundary padding method, the built-in Matlab function "symmetric"

was used and that the results were smoothed by a Gaussian filter ($\sigma = 0.5$ degrees). The average value of pixels inside an object boundary of the edge-content information map (shown in Figure 1c) was used to represent that object's edge-content information.

To derive matching control objects for all text objects, non-text objects were binned by eccentricity ($<$ 200, between 200 and 300, and $>$ 300 pixels) and size ($<$ 1,650, between 1,650 and 5,600, and $>$ 5,600 pixels). 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 identical eccentricity (Ecc.) and similar saliency (Sal.), luminance contrast (LumC.), and edge-content information (EdgeC.). 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.

In order to measure the attraction of visual attention, object-based eye movement measures were used. We used one major measure, *fixation probability* (the probability of a fixation to land inside a text or non-text object or a control region during a trial), and two minor measures, *minimum fixation distance* (the shortest Euclidean distance from the center of the object or region to any fixation during a trial) and *first acquisition time* (the time from stimulus presentation to first target fixation). In every analysis, the major measure was used first in order to examine fixation preference, and subsequently the minor measures were used to support the major measure or to detect any effects when the major measure did not reveal any differences. One drawback of the fixation probability measure is that when there is no fixation landing inside an object boundary, the fixation probability for that object is 0 regardless of how closely a fixation approached it. The same drawback exists for first acquisition time; it may not be representative when fixation probability is low and only few data points become available. Minimum fixation distance was computed to overcome this drawback and provide convergent evidence for any attractiveness results. According to Nuthmann and Henderson (2010), viewers have a tendency to saccade to the center of objects in order to examine them. Their result may support the psychological validity of the measure of minimum fixation distance proposed in this study. Higher fixation probability, shorter first acquisition time, and shorter minimum fixation distance were considered to indicate stronger attraction of attention by a given object. A within-subject one-way analysis of variance (ANOVA) was used to examine the main effect of object category (texts vs. non-texts vs. control regions), and then Bonferroni corrected post-hoc tests were used for the comparison of conditions.

	Size	Ecc.	Sal.	LumC.	EdgeC.
Experiment 1					
Text	2,631 (2.92)	283 (9.43)	0.39	40	0.65
Non-text	2,828 (3.14)	292 (9.73)	0.40	40	0.64
Con. region	2,631 (2.92)	283 (9.43)	0.35	46	0.53
Experiment 2					
Erased text	2,631 (2.92)	283 (9.43)	0.41	21	0.48
Non-text	2,676 (2.97)	293 (9.77)	0.41	24	0.57
Con. region	2,631 (2.92)	283 (9.43)	0.35	36	0.45
Experiment 3					
UncText H B	2,351 (2.61)	288 (9.60)	0.20	10	0.22
UncText INH B	2,723 (3.03)	281 (9.37)	0.36	55	0.59
UncText H	2,351 (2.61)	288 (9.60)	0.25	34	0.43
UncText INH	2,723 (3.03)	281 (9.37)	0.36	57	0.69
Non-Text H	2,670 (2.97)	301 (10.03)	0.28	34	0.53
Non-Text INH	2,746 (3.05)	284 (9.47)	0.38	57	0.69
Con. region H	2,351 (2.61)	287 (9.57)	0.26	40	0.50
Con. region INH	2,723 (3.03)	281 (9.37)	0.37	56	0.61

Table 1. Average characteristics of text objects, non-text objects, and control regions. Size and eccentricity (Ecc.) are shown in pixels, and degrees of visual angle are shown in parentheses. Furthermore, saliency (Sal.), luminance contrast (LumC.), and edge-content information (EdgeC.) are presented. *Notes*: H, texts in front of homogenous background; INH, texts in front of inhomogenous background.

Results and discussion

Fixation probability and minimum fixation distance of texts, non-texts and control regions are shown in Figure 2. The main effect of object category (texts vs. non-texts vs. control regions) on fixation probability was significant, $F(2; 28) = 98.26$, $p < 0.001$. Post-hoc tests revealed that the fixation probability of texts ($M = 0.18$, $SD = 0.05$) was significantly higher than the one of non-text objects ($M = 0.08$, $SD = 0.02$) and control regions ($M = 0.03$, $SD = 0.01$), both $ps < 0.001$. Furthermore, non-text objects had higher fixation probability than control regions, $p < 0.001$, which may be due to control regions not having an obvious boundary like text and non-text objects. This result is in line with the finding of Nuthmann and Henderson (2010) that viewers tend to fixate close to the center of objects (and therefore receive higher fixation probability), but not necessarily close to the centers of salient regions that do not overlap with real objects. In terms of the number of text objects in an image, we found that fixation probability decreases as their number increases, $F(2; 42) = 25.52$, $p < 0.001$, when all cases were categorized into bins of 1 to 4 ($M = 0.25$, $SD = 0.07$), 5 to 8 ($M = 0.17$, $SD = 0.07$), and more than 8 text objects ($M = 0.09$, $SD = 0.03$) with roughly the same number of cases in each bin. Post-hoc analysis indicated that all groups differed significantly, $ps < 0.01$. The results may be due to multiple text objects competing with each other, and the 3-second viewing may be insufficient for viewers to explore all text objects. Since we set up the same number of control regions for text objects in the same images, the number of text objects in an image should not influence the overall results.

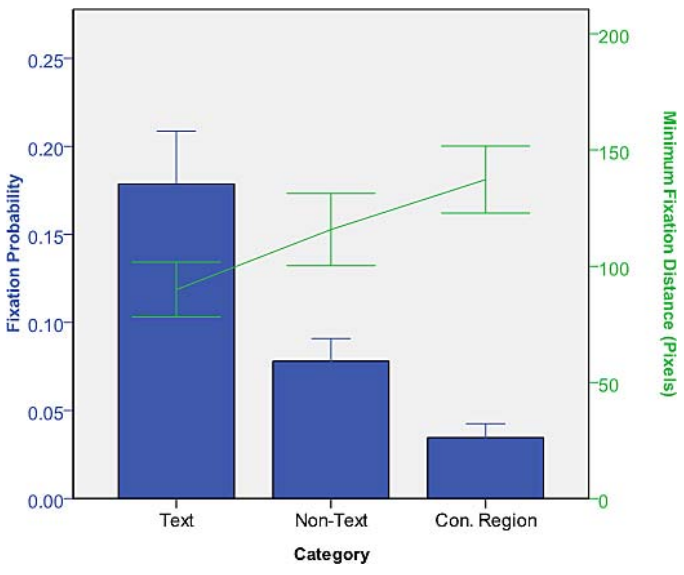


Figure 2. Fixation probability and minimum fixation distance of texts, non-texts, and control regions in Experiment 1. In this chart and all following ones, error bars are based on 95% confidence intervals.

We used minimum fixation distance instead of first acquisition time for additional analysis because average fixation probability was low (< 0.2). The main effect of object category on minimum fixation distance was significant $F(2; 28) = 106.06$, $p < 0.001$. Minimum fixation distance was shorter for texts ($M = 89.93$, $SD = 21.36$) than for non-text objects ($M = 115.79$, $SD = 28.05$) and control regions ($M = 137.31$, $SD = 26.03$), $ps < 0.001$. Furthermore, non-text objects had shorter minimum fixation distance than control regions, $p < 0.001$. In summary, the consistency of these results suggests that texts were more attractive than both non-text objects and control regions.

The selected controls attempted to separate the contribution of low-level salience from high-level features such as expected locations, dependencies among objects or global statistics from the entire image, or unique visual features of texts to the allocation of visual attention. Texts, like faces, might have unique visual features that are unrelated to typical low-level visual saliency. Human observers may have developed "text detectors" during everyday scene viewing that are sensitive to these features and guide attention toward them. We will test how expected locations of texts affect eye movements in Experiment 2, and the potential influence of unique visual features of texts on attention will be examined in Experiment 3.

Experiment 2: erased text

To test whether the typical 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.

Method

Participants

Fifteen participants performed this experiment. All were students at the University of Massachusetts Boston, aged 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 (SR Research, Osgoode, ON, Canada) 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

(Dell Inc., Round Rock, TX) with a refresh rate of 85 Hz and a screen resolution of 1,024 × 768 pixels.

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, using the Adobe Photoshop 9.0 software (Adobe Inc., San Jose, CA), 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 (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 (Itti, Koch, & Niebur, 1998). Note that luminance contrast was computed exclusively within an object, but saliency was calculated according to the whole image, and the neighboring pixels 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 while its luminance contrast is reduced since there is minimal contrast inside the sign.

Procedure

After participants read the instructions, a standard nine-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 scenes. 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 10-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.

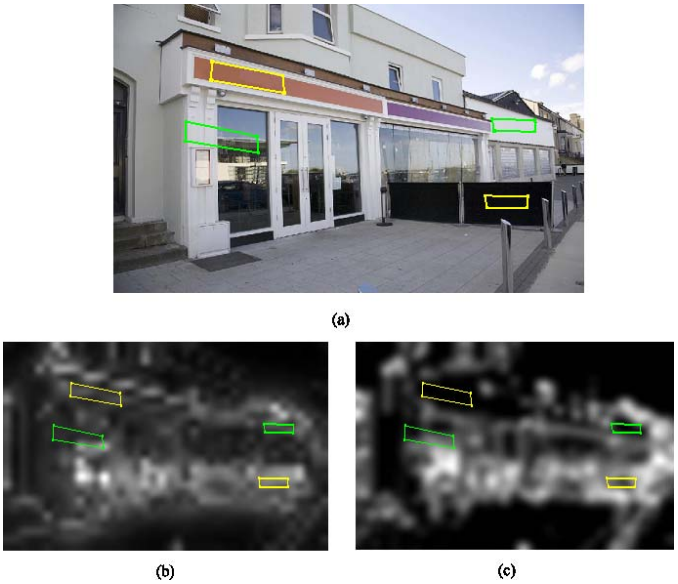


Figure 3. (a) Erased texts (yellow polygons) and their paired control regions (green polygons) in a sample stimulus for Experiment 2. The corresponding saliency and edge-content information are illustrated in (b) and (c). Note that the saliency and edge-content information of erased texts regions were reduced compared to Figure 1, and therefore the control regions were chosen differently.

Analysis

The raw eye-movement data were processed using the standard EyeLink parser (EyeLink User Manual v. 1.4.0, SR Research). To investigate the attractiveness of texts during the initial visual scanning of the scenes, eye fixation data were only analyzed for the first 3 seconds of the viewing duration.[1] In the same manner as performed in Experiment 1, non-text objects and control regions were chosen based on similar size, eccentricity, saliency, and luminance contrast (Table 1). As mentioned above, the luminance contrast within the regions of removed texts was low due to these regions being "plain" after the text removal, but the saliency was affected less. For control regions, we were not able to match both saliency and luminance contrast, since these two variables were positively correlated, $r = 0.34$, for a randomly selected region from the given eccentricity. The luminance contrast of control regions (36) was higher than that of removed-text regions (21). We will further discuss this in the following section.

Results and discussion

The main effect of object category (erased text vs. non-text vs. control region) on fixation probability was significant, $F(2; 28) = 17.02$, $p < 0.001$, as shown by a within-subject one-way ANOVA (Figure 4). Post-hoc tests revealed that while erased texts ($M = 0.07$, $SD = 0.02$) had slightly higher fixation probability than non-text objects ($M = 0.06$, $SD = 0.02$), this difference was not statistically significant, $p = 1.00$. Both erased text and non-text objects received higher fixation probability than control regions ($M = 0.03$, $SD = 0.01$), both $ps < 0.01$.

For additional analysis, minimum fixation distance was used because average fixation probability was low (< 0.1). The main effect of object category on minimum fixation distance was significant, $F(2; 42) = 8.27$, $p < 0.01$. A post-hoc test indicated that minimum fixation distance for erased texts was shorter than for

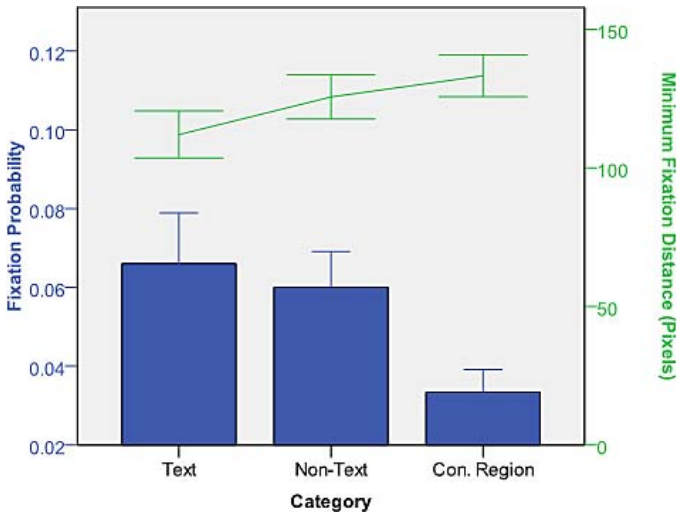


Figure 4. Fixation probability and minimum fixation distance of texts, non-texts, and control regions in Experiment 2.

non-text objects, $t(14) = 5.06$, $p < 0.001$ and for control regions, $t(14) = 8.40$, $p < 0.001$. Furthermore, minimum fixation distance for non-text objects was shorter than for control regions, $t(14) = 2.35$, $p < 0.05$. These results show that viewers did not fixate inside the boundaries of typical locations of text, which may be due to the plainness caused by text removal. However, the results of minimum fixation distance indicated that viewers paid a disproportionate amount of attention to the text removal regions within the scene.

The findings of Experiment 2 indicate that part of the attractiveness of texts derives from their prominent, expected locations in typical real-world images. This effect might be caused by dependencies among objects or global statistics within the entire scene. For example, viewers might recognize a store banner from its positions within the building layout, and they might be attracted by this banner region even without texts. However, Einhäuser and König (2003) pointed out that strong local reductions of luminance-contrast attract fixations. We consider this factor part of saliency because we found that the text removal regions still carried high saliency although their luminance contrasts were strongly reduced. We tried to match saliency between text removal regions and controls as much as possible in order to separate the contribution of low-level saliency from high-level features (i.e., expected location and special features of texts) to fixation positions.

Experiment 3: unconstrained text

To eliminate the influence of expected locations and test whether the unique visual features of text by themselves attract attention, Experiment 3 dissociated texts from their typical locations and placed them in front of homogeneous or inhomogeneous backgrounds. The purpose of using inhomogeneous backgrounds was to add visual noise (non-text patterns) to the unique visual features of text (text pattern), and we expected to find less attraction of attention by texts in front of such inhomogeneous backgrounds.

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. All were students aged 19 to 40 years old and had normal or corrected-to-normal vision. Ten dollars were received by each participant for a half-hour session.

Apparatus

Eye movements were recorded using an EyeLink Remote system (SR Research) with a sampling frequency of 1000 Hz. Subjects sat 65 cm from an LCD monitor. A chin rest was provided to minimize head movements. The spatial accuracy of the system is about 0.5 degrees of visual angle. Although viewing was binocular, eye movements were recorded from the right eye only. Other settings were the same as in 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 was 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 (Figure 5). 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., in 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 pixel windows covering the candidate region.

As discussed, inhomogeneous backgrounds might cause visual noise that interferes with the unique visual

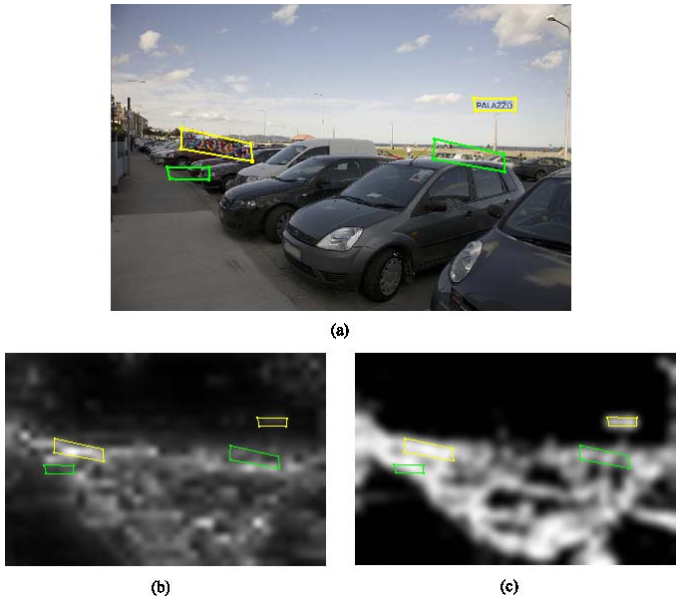


Figure 5. (a) Unconstrained texts (yellow polygons) placed in front of homogeneous (right) and inhomogeneous backgrounds (left) and their paired control regions (green polygons) in one of the scene stimuli. The corresponding saliency and edge-content information are illustrated in (b) and (c).

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 those of the unconstrained texts in front of inhomogeneous background before (UncText INH B) and after (UncText INH) the text parts were placed.

Procedure

The procedure was identical to Experiment 2.

Analysis

The analyses were identical to Experiment 2. Three-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.

Results and discussion

For fixation probability, a within-subject two-way ANOVA showed that the main effect of object category (texts vs. non-texts vs. control regions) was significant, $F(2; 28) = 37.53$, $p < 0.001$, the main effect of background (homogeneous vs. inhomogeneous) was also significant, $F(1; 14) = 4.70$, $p < 0.05$, and the interaction of object category and background was significant as well, $F(2; 28) = 24.87$, $p < 0.001$. As illustrated in Figure 6a, this interaction can be explained by the object category effect being more pronounced for homogeneous than for inhomogeneous background. A within-subject one-way ANOVA revealed that the main effect of object category for homogeneous background was significant, $F(2; 28) = 38.68$, $p < 0.001$. The fixation probability of unconstrained texts in front of homogeneous background ($M = 0.18$, $SD = 0.09$) was higher than for non-texts ($M = 0.05$, $SD = 0.02$) and control regions ($M = 0.02$, $SD = 0.01$), both $ps < 0.001$. The main effect of object category for inhomogeneous background was significant as well, $F(2; 28) = 19.37$, $p < 0.001$. The fixation probability for texts ($M = 0.11$, $SD = 0.05$) was still significantly higher than for non-texts ($M = 0.06$, $SD = 0.03$) and control regions ($M = 0.04$, $SD = 0.02$), $ps < 0.01$, but the difference was not as large as for texts in front of homogeneous background.

For minimum fixation distance, a corresponding within-subject two-way ANOVA also revealed significant main effects of object category, $F(2; 28) = 10.79$, $p < .001$, and background, $F(1; 14) = 18.07$, $p < 0.01$, and their interaction was also significant, $F(2; 28) = 11.77$, $p < .001$. Within-subject one-way ANOVAs showed a significant main effect for homogeneous background, $F(2; 28) = 12.36$, $p < 0.001$, and for inhomogeneous background, $F(2; 28) = 3.56$, $p < 0.05$. The post-hoc tests revealed that for homogeneous backgrounds, minimum fixation distance was significantly higher for unconstrained texts ($M = 120.48$, $SD = 34.16$) than for non-text objects ($M = 139.64$, $SD = 23.21$) and control regions ($M = 147.29$, $SD = 22.51$), $ps < 0.05$. For inhomogeneous background, minimum fixation distance of unconstrained texts ($M = 128.12$, $SD = 26.49$) was significantly higher than the one of control regions ($M = 134.22$, $SD = 22.38$), $p < 0.05$. As shown in Figure 6b, the trends were similar to fixation probability; unconstrained texts in front of homogeneous and inhomogeneous background received shorter distances than did control objects and regions and can therefore be considered more attractive.

To summarize, we found texts in front of homogeneous background (Text H) to be more attractive than texts in front of inhomogeneous background (Text INH; Figures 6a and 6b). Regions with higher low-level saliency measures tend to receive more attention, but the opposite result was observed, i.e., Text INH was associated with higher saliency, luminance contrast, and edge-content information than Text H (Table 1), but received less attention. Therefore, our data imply

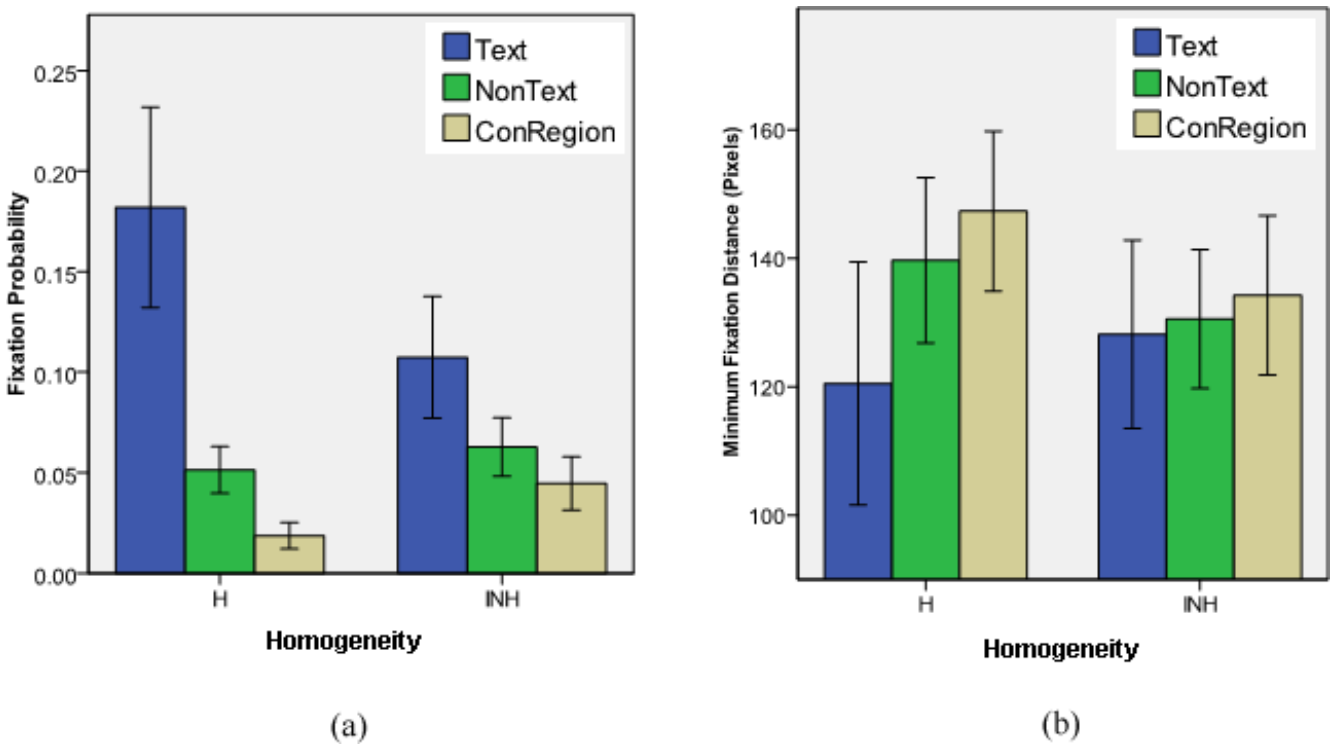


Figure 6. Fixation probability (a) and minimum fixation distance (b) of unconstrained texts in front of homogeneous (H) and inhomogeneous (INH) background, and the corresponding values for non-text objects and control regions.

that the distinctive visual features of texts might be superior to low-level saliency measures in attracting attention.

It should be noted that participants being attracted by texts and actually "reading" texts are two different matters, and this study focused on how participants' attention was caught by texts. Text INH containing both text and non-text patterns may or may not be "recognized" as text due to the noise level and position, but they did draw more attention than controls (Figure 6b).

Furthermore, it must be pointed out that the unconstrained texts could be considered as object-scene inconsistencies (specifically, syntactic violations and maybe semantic violations) since they were placed in unexpected locations in other scenes. Scene inconsistencies have been a highly debated issue, and previous studies either found them to influence initial eye movements (e.g., Loftus & Mackworth, 1978; Becker, Pashler, & Lubin, 2007; Bonitz & Gordon, 2008; Underwood & Foulsham, 2006; Underwood, Humphreys, & Cross, 2007; Underwood, Templeman, Lamming, & Foulsham, 2008) or failed to obtain evidence for such early detection (e.g., Gareze & Findlay, 2007; Rayner, Castelhano, & Yang, 2009; Võ & Henderson, 2009, 2011).

Regardless of this debate, it is clear that at least in some instances, a text placed in an unexpected location, e.g., floating in mid-air, captures attention, which may be due to its specific visual features or its unusual placement. The latter case would also apply to any non-text object placed in the same way. To resolve the potential issue of unusual placement of texts that arose in this experiment, in Experiment 4 we placed both texts and line drawings of the objects described by the texts in unexpected locations.

Experiment 4: unconstrained texts and line drawings

We placed an item-pair—a text and a drawing—in unexpected locations in a scene. If the text were found to attract more attention than the drawings, it would confirm the contribution of specific visual features of texts to their attractiveness. Texts and drawings were placed either in front of homogeneous or inhomogeneous *backgrounds*. We expected to observe similar results to the ones found in Experiment 3, that is, the attraction of visual features of texts being degraded by noise. In addition to comparing texts and drawings, we compared two *text-types*, namely texts (regular words) and their scrambled versions (i.e., all letters of the word being randomly rearranged in such a way that they did not form another English word), in order to test if higher-level processing, such as semantics, influences the attraction of attention.

Method

Participants

Twelve students from the University of Massachusetts at Boston participated. All were students with normal or corrected-to-normal vision and aged 19 to 40 years old. Each participant received 10 dollars for a half-hour session.

Apparatus

The apparatus was the same as in Experiment 3.

Stimuli

Two hundred new natural-scene images, which were not used in Experiments 1 to 3, were selected from the LabelMe dataset. Eighty of these images were randomly selected to be superimposed with one item-pair each. The other 120 images were presented without any modification. There were four versions of the 80 superimposed images, resulting in 320 images for a counterbalanced design (i.e., one viewer only saw one of the 4 versions of the stimuli). Each observer viewed 80 item-pairs (cases). Figure 7 shows an example of all four versions of the same stimulus with items drawn on homogeneous background. For the placement of texts and line drawings, two different items (items A and B) were chosen for each scene, and their addition to the scene was performed in four different versions: either (a) a word describing item A (e.g., "sled" as shown in Table 2) and a drawing of item B, (b) a word describing item B (e.g., "yoyo") and a drawing of item A, (c) a scrambled version of a word describing item A (e.g., "dsle") and a drawing of item B, and (d) a scrambled version of a word describing item B (e.g., "yyoo") and a drawing of item A. The length of regular and scrambled words ranged between 3 and 11 letters (average: six letters). The eccentricity of the text or the drawing was randomly assigned and varied between 200 and 320 pixels (average: 253 pixels). The minimum polar angle, measured from the screen center, between the text and the drawing in each image was set to 60° to avoid crowding of the artificial items. All texts and drawings were resized to cover approximately 2,600 pixels. Table 3 shows the characteristics of texts and drawings in front of homogeneous (H) and inhomogeneous backgrounds (INH).

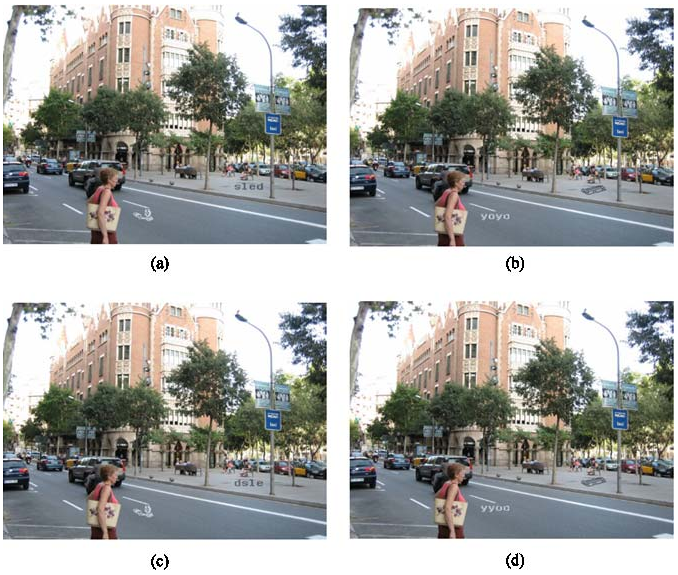


Figure 7. An example of the four stimulus versions of stimuli used in Experiment 4, with words and drawings on homogeneous background. (A) Word of Item A (sled) vs. drawing of Item B, (b) word of Item B (yoyo) vs. drawing of Item A, (c) scrambled word of Item A (dsle) vs. drawing of Item B, and (d) scrambled word of Item B (yyoo) vs. drawing of Item A.

Procedure

Equal numbers of subjects viewed stimuli from conditions a, b, c, and d in a counter-balanced design (described above), and each stimulus was presented for 5 seconds. The software "Eyetrack" developed by Jeffrey D. Kinsey, David J. Stracuzzi, and Chuck Clifton, University of Massachusetts Amherst, was used for recording eye movements. This software provides an easy-to-use interface for between-subject designs and has been widely used in the community of eye-movement researchers. Other settings were identical to Experiments 2 and 3.

Analysis

Fixation probability, minimum fixation distance, and first acquisition time were examined using a within-subject three-way ANOVA including item-type (texts vs. drawings), text-type (regular vs. scrambled), and background (homogeneous vs. inhomogeneous). There were 20 cases per condition. The fixation probability ANOVA served as the main analysis, while the ANOVAs for minimum fixation distance and first acquisition time were considered additional analyses. One participant was excluded from the analysis of first acquisition time since his fixation probability of drawings was 0 in one condition.

Results and discussion

For fixation probability, the main effects of item-type and text-type did not reach significance, all Fs(1; 11) < 2.48, $ps > 0.1$, but the main effect of background

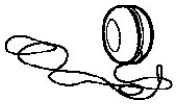
	Item A	Item B
Word (scrambled word)	sled (dsle)	yoyo (yyoo)
Object drawing		

Table 2. Examples of texts (words and scrambled words) and object drawings used in Experiment 4.

did, $F(1; 11) = 83.85$, $p < 0.001$. Fixation probability was higher in front of homogeneous background than inhomogeneous background. All interactions among item-type, text-type, and background failed to reach significance, all $Fs(1; 11) < 0.59$, $ps > 0.46$. These results suggest that both texts and drawings drew more attention when they were presented on a clear background than when they were degraded by an inhomogeneous background.

For minimum fixation distance, a three-way ANOVA yielded main effects for item-type and background, both $Fs(1; 11) > 33.17$, $ps < 0.001$, but not for text-type, $F(1; 11) = 0.35$, $p = 0.57$. All interactions among item-type, text-type, and background were non-significant, $Fs(1; 11) < 3.08$, $ps > 0.11$. Minimum fixation distance was shorter for texts than drawings, and it was also shorter for homogeneous background than inhomogeneous background.

The results of the first acquisition time again demonstrated significant main effects of item-type and background, both $Fs(1; 10) > 13.96$, $ps < 0.01$, but not for text-type $F(1; 10) = 1.42$, $p = 0.26$. The interactions among item-type, text-type, and background were not significant, $Fs(1; 10) < 2.56$, $p > 0.14$, except for a marginal interaction between item-type and text-type, $F(1; 10) = 3.60$, $p = 0.09$. Surprisingly, items in front of inhomogeneous background seemed to receive fixations earlier than those in front of homogeneous background. It should be noted, however, that first acquisition time only accounted for items being fixated. When the background was homogeneous, the average fixation probability was more than 0.55. In contrast, the average fixation probability was only approximately 0.35 when items were in front of inhomogeneous background. Here we analyze first acquisition time separately for homogeneous and inhomogeneous background because the fixation probabilities in these conditions were incompatible. For homogeneous background, a two-way ANOVA yielded a significant main effect of item-type, $F(1; 10) = 7.61$, $p < 0.05$, but not for text-type nor the interaction, both $Fs(1; 10) < 2.50$, $ps > 0.15$. For inhomogeneous background, there were no significant main effects of item-type and text-type, nor a significant interaction, all $Fs(1; 10) < 0.24$, $p > 0.62$. The results indicated that first acquisition time was shorter for texts than for drawings when the background was homogeneous, but no effect was found for inhomogeneous background. The averages and standard deviations of fixation probability, minimum fixation distance, and first acquisition time are shown in Figure 8.

The results of minimum fixation distance and first acquisition time were consistent with regard to texts receiving more attention than drawings, suggesting that the specific visual features of texts cause their attractiveness advantage. By definition, the scrambled words in Experiment 4 were not dictionary words, but it is important to note that their word length was controlled compared to their paired (regular) words. We did not find statistical differences between words and scrambled words in any of the measures, $Fs(1; 11) < 2.48$, $ps > 0.1$. These data suggest that the attention-capturing features of texts are operating at a low level so that the attraction of attention does not seem to depend on whether a word carries meaning.

The results of Experiment 4 confirmed that texts are more attractive than non-texts. Both words and scrambled words were found more attractive than line drawings depicting the corresponding objects. Because words and scrambled words yielded similar attractiveness results, the attraction of attention by texts seems to be caused by low-level visual features, not high-level semantics. This result raises important questions: Are these low-level features, such as the regular spacing and similarity of characters, specific to the observer's native writing system? Does a simple image transformation

	Size	Ecc.	Sal.	LumC.	EdgeC.
H					
Texts	2,699 (3.00)	262 (8.73)	0.21	36.75	0.66
Drawings	2,652 (2.95)	262 (8.73)	0.23	38.26	0.64
INH					
Texts	2,700 (3.00)	258 (8.60)	0.32	51.64	0.78
Drawings	2,652 (2.95)	258 (8.60)	0.33	52.15	0.79

Table 3. Average characteristics of texts and drawings in Experiment 4. *Notes*: H, texts in front of homogenous background; INH, texts in front of inhomogenous background.

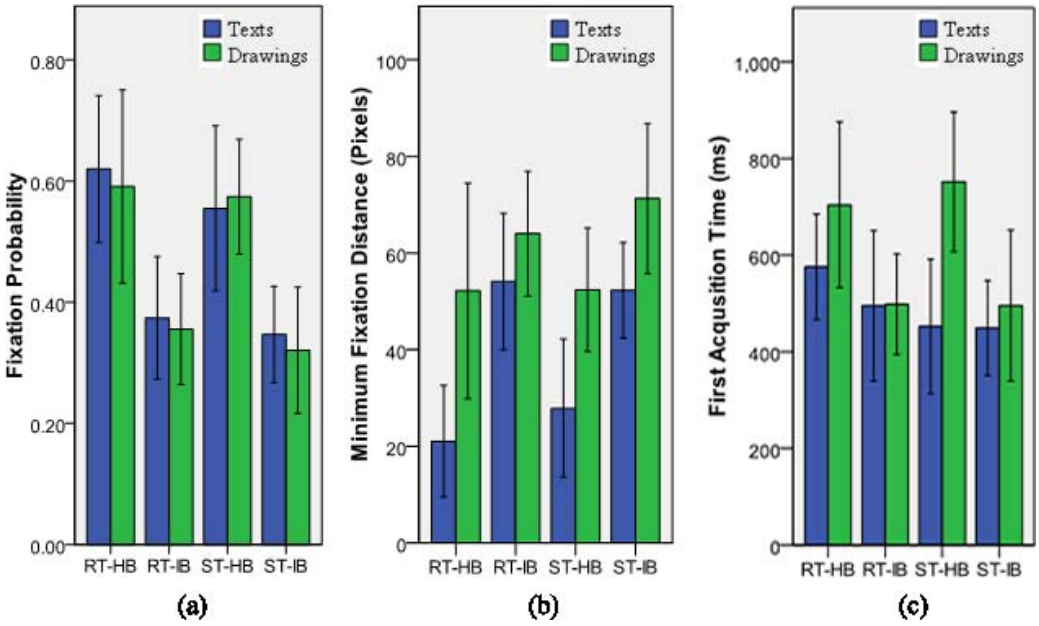


Figure 8. Results of Experiment 4 for texts and drawings. (a) Fixation probability, (b) minimum fixation distance, and (c) first acquisition time (RT: regular text, ST: scrambled text, HB: homogeneous background, and IB: inhomogeneous background).

such as rotation by 180° preserve their attractiveness? These questions were addressed in Experiment 5.

Experiment 5: upside-down English and Chinese texts

To study the influence of the observers' familiarity with their native writing system, we carried out a further experiment by placing texts in Experiment 1 upside-down or replacing them with Chinese texts. These stimuli were presented to English speakers. The rationale for using upside-down English texts was to keep the low-level features such as regular spacing and similarity of letters but reduce possible influences of higher-level processing such as meaning. Chinese texts were chosen because they are visually dissimilar to texts in the English language and other alphabetic writing systems. Our hypothesis is that subjects may have developed specific "text detectors" for their native writing system during everyday life so that their attention would be biased toward words in that writing system.

After the conclusion of this experiment, we also received an opportunity to test native Chinese speakers. Since we found that turning texts upside-down did not affect attentional capture for English speakers, we decided to use exactly the same materials for the Chinese subjects without turning the Chinese texts upside-down for better comparability of results between the subject groups.

Method

Participants

In the group of non-Chinese English speakers, an additional 14 students from the University of Massachusetts at Boston participated in this experiment. All of them were native speakers of English, and none of them had learned any Chinese or had participated in Experiments 1 to 4. For the group of Chinese speakers, 16 native speakers of Chinese were recruited at China Medical University, Taiwan. Each participant received 10 US dollars or 100 Taiwan dollars, respectively, for participation in a half-hour session. All had normal or corrected-to-normal vision.

Apparatus

Eye movements were recorded using EyeLink 1000 Remote systems (SR Research) both at the University of Massachusetts at Boston and at China Medical University, Taiwan. Other settings were the same as in Experiments 2 and 3.

Stimuli

As shown in Figure 9, the original texts from Experiment 1 were either rotated by 180° or replaced by Chinese texts. Figure 9a illustrates C1, in which half of the original texts were rotated and the other half was replaced with Chinese texts. In C2, as demonstrated in Figure 9b, the upside-down texts in C1 were replaced with Chinese texts, and the Chinese texts in C1 were replaced with the original, but upside-down, English

(a) (b)

Figure 9. Example of upside-down and Chinese texts used in Experiment 5. (a) Version C1, in which half of the original texts were rotated and the other half was replaced with Chinese texts. (b) Version C2, in which the upside-down texts in C1 were replaced with Chinese texts, and the Chinese texts in C1 were replaced with upside-down texts.

texts. Table 4 shows the characteristics of the upside-down and Chinese texts in C1 and C2. The characteristics of all upside-down and Chinese texts in C1 and C2 were very similar to those of the original texts in Experiment 1.

Procedure

The procedure was identical to Experiments 2 and 3 except that half of the subjects viewed condition 1 (C1) stimuli and the others viewed condition 2 (C2) stimuli in a between-subject counter-balanced design (described below). The same Eyetrack software as in Experiment 4 was used for recording eye movements.

Analysis

The analyses were identical to Experiments 2 and 3. Similar to Experiments 1 to 4, three-second viewing durations were analyzed for each trial. For English speakers, 7 subjects viewed C1 and 7 subjects viewed C2, and those data were combined so that upside-down English text and Chinese text for each item were viewed in a between-subject counter-balanced design. The same analysis was performed for Chinese speakers.

Results and discussion

For English speakers, as shown in Figure 10, fixation probability was higher for upside-down texts than for Chinese texts, $t(13) = 5.62$, $p < 0.001$. This result suggests that upside-down English texts attract English speakers' attention more strongly than Chinese texts do. This trend is consistent with the results of minimum fixation distance, which was slightly shorter for upside-down texts (83.69) than for Chinese texts (88.16), but the difference failed to reach significance level, $t(13) = 1.63$, $p > 0.1$. A between-experiment comparison revealed that turning texts upside-down did not lead to any changes in their attraction of attention (see General Discussion for between-experiment analyses).

For Chinese speakers, the results were reversed as compared to English speakers; fixation probability was lower for upside-down English texts than for Chinese texts, $t(15) = 3.67$, $p < 0.01$. Minimum fixation distance was shorter for Chinese texts than for upside-down English texts, $t(15) = 4.46$, $p < 0.01$.

In the comparison between English and Chinese speakers, we found that Chinese texts were fixated equally often by both groups, but the upside-down texts were fixated more often by English speakers than by Chinese speakers. In other words, only the English speakers were biased toward their own native language. One possibility is that other factors played a role, such as expected locations, e.g., Chinese speakers might expect texts on vertical rather than horizontal signs given that most stimulus images were taken in North America and Europe. Nevertheless, based on the results of English speakers, Experiment 5 suggests that attraction of attention depends to some extent on the observer's familiarity with the writing system and language. The reason might be that English viewers have developed stronger "text detectors" for English texts during everyday life. The results may support the implication suggested in Cerf et al. (2009) that the allocation of attention to text is developed through learning.

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

Experiment 5	Size	Ecc.	Sal.	LumC.
Upside-down text C1	2,227 (2.47)	273 (9.10)	0.43	38
Chinese text C2	2,255 (2.50)	273 (9.10)	0.42	37
Upside-down text C2	3,003 (3.34)	292 (9.73)	0.40	38
Chinese text C1	2,996 (3.33)	292 (9.73)	0.39	37

Table 4. Average characteristics of upside-down and Chinese texts in each condition.

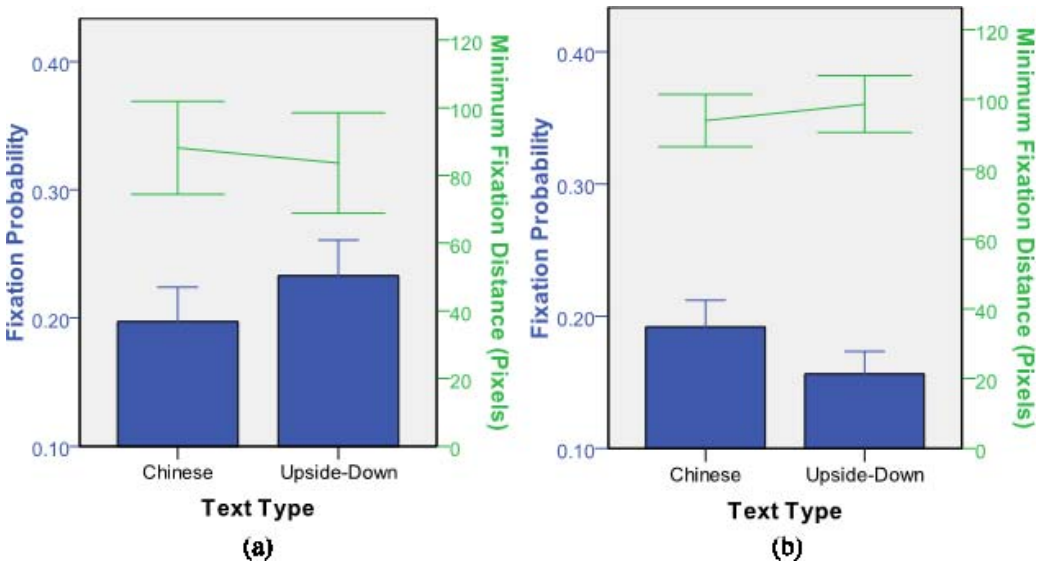


Figure 10. Fixation probability and minimum fixation distance of Chinese and upside-down English texts for (a) English readers and (b) Chinese readers.

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 be assumed to 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" modeled by Torralba et al. (2006) and the influence of expected locations on visual attention as pointed out by Eckstein 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 features typically associated with saliency in their ability to attract attention, and their influence on attention was reduced by the noise caused by inhomogeneous background. However, the results obtained in Experiment 3 might also have been caused by the replacement of texts inducing oddness through semantic or syntactic violation. Experiment 4 provided convergent evidence for the contribution of the specific visual features to text attractiveness by placing texts and object drawings in unexpected locations and still finding stronger attentional capture by texts. In addition, Experiment 4 indicated that this capture might be caused by low-level visual features rather than high-level semantics since words and scrambled words yielded similar results. Experiment 5 further investigated how familiarity influences the attraction of attention by texts by presenting upside-down English and upright Chinese texts to native English and Chinese speakers. The results showed that viewers were biased toward their native language, which indicates that familiarity affects the allocation of attention. We conclude that both low-level specific visual features of texts and, to a lesser extent, high-level features (expected locations) contribute to the ability of texts to attract a disproportionate amount of visual attention in real-world scenes.

The results obtained from Experiment 1 might serve as a starting point for other experiments. In Experiment 2, fixation probability for erased texts (mean: 0.07) dropped in comparison to text objects in Experiment 1 (mean: 0.18), $F(1; 28) = 35.82$, $p < 0.001$, for a between-subject ANOVA. Minimum fixation distance was significantly longer for erased texts in Experiment 2 (mean: 111.98) than for texts in Experiment 1 (mean: 89.93), $F(1; 28) = 10.53$, $p < 0.01$. This result might be caused by the reduction of saliency and luminance contrast that accompanied the erasure of text. In Experiment 3, Fixation probability of the unconstrained texts in front of homogeneous background was not statistically different from that of texts in Experiment 1 located in expected positions (both means: 0.18), $F(1; 14) = 0.01$, $p > 0.9$. This finding suggests that the specific text features might cause stronger attraction than expected locations. For Experiment 5, it is interesting to point out that the fixation probability of viewers' non-native language stimuli was considerably high (0.20 for upside-down texts viewed by Chinese speaker and 0.16 for upside-down texts viewed by Chinese speaker) compared to the text objects in Experiment 1 (0.18). This finding might imply that there are cross-language features of

texts that capture attention, regardless whether the texts carry meaning. Moreover, turning English texts upside-down does not seem to significantly reduce their capture of English speakers' attention, which provides further evidence for the dominance of low-level factors in attracting attention to texts. However, those implications from between-experiment comparisons need to be verified in further well-controlled experiments, for example, an experiment containing regular, erased, upside-down texts in a between-subject counterbalanced design. Furthermore, to follow up on Experiment 2, another experiment could be conducted by erasing non-text regions by filling them with a background color, and then comparing them in terms of their attentional capture to text-removal regions. Both cases in such design cause strong reduction of luminance contrasts, but only text-removal regions occupy expected locations for texts. Such investigations will be pursued in future studies.

The free viewing task seems to be less constrained as compared to visual search or memorization tasks. Search and memorization tasks require specific top-down control of attention that might dominate task performance and therefore lead to different results from those obtained in the present study. However, during free viewing tasks, observers might attempt to retrieve as much information as possible, including deliberately looking for texts in order to make the scene more interpretable and contribute to its understanding and memorization. Therefore, although the task was free viewing and we included text-absent images in all experiments, we cannot rule out the possibility that observers may actually perform text searching and memorizing.

It would be interesting to see how texts are "read" in real-world scenes. In our previous study (Wang, Hwang, & Pomplun, 2010), fixation durations were found to be influenced by object size, frequency, and predictability, and we suggested that the recognition of objects in scene viewing shares some characteristics with the recognition of words in reading. It is important to analyze the underlying factors affecting processing time of texts in real-world scenes and compare the results to existing text reading studies (Rayner, 2009).

There are other factors, i.e., scene context and scene syntax, which might affect expected locations. For instance, Torralba et al. (2006) developed a computational model of "contextual guidance" according to global scene statistics. Furthermore, Hwang, Wang, and Pomplun (2011) proposed "semantic guidance" during scene viewing which leads to a tendency toward gaze transitions between semantically similar objects in the scene. It was also found that "object dependency" (i.e., the statistical contingencies between objects, such as between a plate and silverware) can help viewers to predict the location of other objects from a given object or scene (Oliva & Torralba, 2007). For a better understanding of the attentional bias toward texts, it may thus be important to further extract the object dependency between texts and other objects from an image dataset such as LabelMe. Using the concepts of contextual guidance, semantic guidance, and object dependency, a computational model for human text detection could be developed.

There are many text-like patterns such as windows, fences, or brick walls that are easily misidentified as texts by artificial text detectors (Ye, Jiao, Huang, & Yu, 2007). Furthermore, in Experiment 5 we found that English and Chinese-speaking viewers possess different preferences for the attraction of their attention to texts. Future research could study the influences of specific visual features of texts to human viewers, using the analysis of eye movements. For example, such experiments could test the contribution of individual features of texts, e.g., orientations or arrangements of letters and strokes, to low-level attraction of human viewers' attention. Furthermore, it might be useful to further investigate the difference of special features between English and Chinese texts, as the results are potentially important for developing more efficient and general text detection algorithms.

Acknowledgments

Parts of the data were presented at the European Conference on Visual Perception (ECVP 2010) and the Asia Pacific Conference of Vision (APCV 2011). Thanks to Melissa Võ, Gregory Zelinsky, and an anonymous reviewer for their helpful comments on earlier versions of the article. Preparation of the article was supported by Grant R01 EY021802 from the National Institutes of Health (NIH) to Marc Pomplun.

Commercial relationships: none.
Corresponding author: Hsueh Cheng-Wang.
Email: hchengwang@gmail.com.
Address: Department of Computer Science, University of Massachusetts at Boston, Boston, MA, USA.

Footnote

[1]The purpose of presenting the scenes for 10 s was to study differences in early versus late scanning. However, we found longer viewing to only increase fixation probabilities and decrease minimum fixation distances without changing the pattern of results.

References

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*(18), 2824–2833.

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 20–30.

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica, 129*(2), 255–263.

Bruce, N. D. B., & Tsotsos, J. K. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems, 18*, 155–162.

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

Cerf, M., Cleary, D., Peters, R., Einhäuser, W., & Koch, C. (2007). Observers are consistent when rating image conspicuity. *Vision Research, 47*, 3052–3060.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real world scenes, saccadic targeting, and Bayesian priors. *Psychological Science, 17*(11), 973–980.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14):18, 1–26, http://www.journalofvision.org/content/8/14/18, doi:10.1167/8.14.18 [PubMed] [Article]

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17*, 1089–1097.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3):3, 1–15, http://www.journalofvision.org/content/8/3/3, doi:10.1167/8.3.3. [PubMed] [Article]

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition (CVPR)*. Anchorage, AK, USA, 1–8.

Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In van Gompel, R., Fischer, M., Murray, W., & Hill, R. W. (Eds.), *Eye movements: A window on mind and brain*. (pp. 537–562). Amsterdam, The Netherlands: Elsevier.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Proceedings of Neural Information Processing Systems (NIPS)*.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498–504.

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In Henderson, J. M. & Ferreira, F. (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. (pp. 1–58). New York: Psychology Press.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision, 9*(5), 1–18, http://www.journalofvision.org/content/9/5/25, doi:10.1167/9.5.25. [PubMed] [Article]

Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research, 51*(10), 1192–1205.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *IEEE International Conference on Computer Vision (ICCV)*. Kyoto, Japan, 2106–2113.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565–572.

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision,* 11(9):9, 1–16, http://www.journalofvision.org/content/11/9/9, doi:10.1167/11.9.9. [PubMed] [Article]

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8):20, 1–19, http://www.journalofvision.org/content/10/8/20, doi:10.1167/8.14.18. [PubMed] [Article]

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Science, 11*(12), 520–527.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual selective attention. *Vision Research, 42*(1), 107–123.

Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, (pp. 1–8). Washington, DC: IEEE Computer Society.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research, 46*(12), 1886–1900.

Rayner, K. (2009). The Thirty Fifth Sir Frederick Bartlett Lecture: Eye movements and attention during reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62*, 1457–1506.

Rayner, K., Castelhano, M. S., & Yang, J. (2009). Eye movements when looking at unusual/weird scenes: Are there cultural differences? *Journal of Experimental Psychology: Learning, Memory & Cognition, 35*, 254–259.

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, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.

Torralba, A., Oliva, A., Castelhano, 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*(4), 766–786.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology, 59*(11), 1931–1949.

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency, and gist in the inspection of objects in natural scenes. In van Gompel, R. P. G., Fischer, M. H., Murray, W. S., & Hill, R. L. (Eds.), *Eye movements: A window on mind and brain*. (pp. 564–579). Amsterdam: Elsevier Science.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition, 17*(1), 159–170.

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision,* 57(2), 137–154.

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, http://www.journalofvision.org/content/9/3/24, doi:10.1167/9.3.24. [PubMed] [Article]

Võ, M. L.-H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics, 73*(6), 1742–1753.

Wang, H.-C., Hwang, A. D., & Pomplun, M. (2010). Object frequency and predictability effects on eye fixation durations in real-world scene viewing. *Journal of Eye Movement Research, 3*(3): 3, 1–10.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Ye, Q., Jiao, J., Huang, J., & Yu, H. (2007). Text detection and restoration in natural scene images. *Journal of Visual Communication and Image Representation, 18*(6), 504–513.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*(4), 787–835.