Visual Attention is Attracted by Text Features Even in Scenes without Text

Hsueh-Cheng Wang (hchengwang@gmail.com)
Shijian Lu (slu@i2r.a-star.edu.sg)
Joo-Hwee Lim (joohwee@i2r.a-star.edu.sg)
Marc Pomplun (marc@cs.umb.edu)
Department of Computer Science, University of Massachusetts at Boston,
100 Morrissey Boulevard, Boston, MA 02125 USA
Institute for Infocomm Research, A*STAR, Singapore
1 Fusionopolis Way, Singapore 138632

Abstract

Previous studies have found that viewers' attention is disproportionately attracted by texts, and one possible reason is that viewers have developed a "text detector" in their visual system to bias their attention toward text features. To verify this hypothesis, we add a text detector module to a visual attention model and test if the inclusion increases the model's ability to predict eye fixation positions, particularly in scenes without any text. A model including text detector, saliency, and center bias is found to predict viewers' eye fixations better than the same model without text detector, even in text-absent images. Furthermore, adding the text detector – which was designed for English texts – improves the prediction of both English- and Chinese-speaking viewers' attention but with a stronger effect for English-speaking viewers. These results support the conclusion that, due to the viewers' everyday reading training, their attention in natural scenes is biased toward text features.

Keywords: real-world scenes; text detector; eye movements; visual attention.

Introduction

When inspecting real-world scenes, human observers continually shift their gaze to retrieve information. Viewers' attention has been found to be biased toward visually salient locations, e.g., high-contrast areas, during scene viewing or search (Itti & Koch, 2001) or toward the center of the screen when viewing scenes on computer monitors (Tatler, 2007). Since it is also known that viewers pay a disproportionate amount of attention to faces (Cerf, Frady, & Koch, 2009), Judd, Ehinger, Durand, and Torralba (2009) equipped their model of visual saliency with a face detector (Viola & Jones, 2004) and a person detector (Felzenszwalb, McAllester, & Ramanan, 2008). In those images that contained depictions of people, their model with all features combined outperformed models trained on typical saliency features such as color, orientation, intensity, and contrast. Cerf et al. (2009) refined the "standard" saliency model by adding a channel of manually-defined regions of faces, texts, and cellphones, and demonstrated that the enhancement of the model significantly improved its ability to predict eye fixations in natural images.

Besides depictions of people, texts in natural scenes are usually important pieces of information, 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. Our previous studies (Wang & Pomplun, 2011; under revision) suggested that attention seems disproportionately attracted by texts but that the specific visual features of texts, e.g., edge density, rather than typically salient features such as color, orientation, intensity, or contrast, are the main attractors of attention. This finding was in line with the results in Baddeley and Tatler (2006) that high spatial frequency edges, not contrasts, predict where we fixate.

Automatic text detection has been a hot topic in the fields of computer vision and pattern recognition for its practical applications. The special features of texts, e.g., the small variation of the stroke width (see Epshtein, Ofek, & Wexler, 2010; Jung, Liu, & Kim, 2009) or edge density (Lu, submitted) have been used to develop text detectors. Although many text detection techniques, i.e., texture-based, region-based, and stroke-based methods, have been reported, many non-text objects, such as windows, fences, or brick walls, easily cause false alarms (see Lu, submitted; Ye, Jiao, Huang, & Yu, 2007, for a review). Furthermore, many established text detectors are restricted under commercial patents. Therefore, only few text detectors are freely available or have been tested in visual attention studies.

Lu, Wang, Lim, and Pomplun (submitted) developed specialized text features, e.g., histograms of edge width and edge density, trained with Support Vector Machine (SVM) classifiers. The study reported better performance compared with earlier studies (e.g., Epshtein, et al., 2010; Jung, et al., 2009) on public text-detecting competition datasets (ICDAR2003 and ICDAR2005). In the present study, we used the automatic text detector developed by Lu et al. (submitted) to test whether it can improve the prediction of viewers' fixations. This detector employs contrast of strokes over background, width of strokes, joints of horizontal and vertical strokes, and stroke structure as key variables

Although manually-defined regions of texts were shown to improve the prediction of eye fixations in text-present images (Cerf et al., 2009), it is unclear if viewers' attention is biased toward any non-text objects which share some features of texts, particularly in text-absent images. In the present study, two eye-movement datasets obtained in our previous investigations (Wang & Pomplun, under revision)

are re-analyzed. The goals of the present study are (1) to investigate the contribution of the automatic text detector to the prediction of eye fixations in real-world scenes, and (2) to verify the hypothesis that viewers' text detection skills are "trained" through exposure to language and affect attentional control even in text-absent scenes

Experiment 1: Unconstrained Texts

We superimposed unconstrained texts onto real-world scenes, i.e., placed them in unexpected locations, in front of either homogeneous background, i.e., in regions with the lowest luminance contrast in the image before placing the text parts, or inhomogeneous background, i.e., those areas with the highest luminance contrast, and found that texts attracted more attention than non-text objects. This dataset is chosen for re-analysis in the present study since the stimuli contain both text-present and text-absent images. Two models, both including saliency and center-bias maps (channels), but one with and one without text-detector map are compared in order to determine whether the inclusion of the text detector improves the prediction of fixations, particularly in text-absent images.

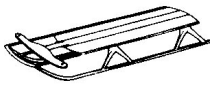
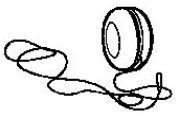
Method

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

Apparatus. Eye movements were recorded using an SR Research EyeLink Remote system with a sampling frequency of 1000 Hz. Subjects sat 65 cm from an LCD monitor approximately 34 x 25 degrees of visual angles. A chin rest was provided to minimize head movements. 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. Although viewing was binocular, eye movements were recorded from the right eye only.

Stimuli. Two hundred natural-scene images were selected from the LabelMe dataset (Russell, Torralba, Murphy & Freeman, 2008). Eighty out of these images were randomly selected to be superimposed with one text and one line drawing. The other 120 images were presented without any modification. For the placement of texts and line drawings, two different items (items A and B in Table 1) were chosen for each scene, and their addition to the scene was performed under four different conditions: either (1) a word describing item A (e.g., "sled" as shown in Table 1) and a drawing of item B, (2) a word describing item B (e.g., "yoyo") and a drawing of item A, (3) a scrambled version of a word describing item A (e.g., "dsle") and a drawing of item B, and (4) a scrambled version of a word describing item B (e.g., "yyoo") and a drawing of item A. All four conditions of text-drawing pairs were presented in a between-subject design, i.e., each participant only viewed one of these conditions. Half of the words (object labels) were placed in front of homogeneous background and the other half were placed on inhomogeneous background. Figure 1 shows an example of all four conditions with words and drawings on homogeneous background. 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 degrees to avoid crowding of the artificial items. All texts and drawings were resized to cover approximately 2500 pixels.

Table 1: Examples of texts (words and scrambled words) and object drawings used in Experiment 1.

	Item A	Item B
Texts	sled (dsle)	yoyo (yyoo)
Object Drawing		

(a) (b) (c) (d)

Figure 1. An example of 4 conditions of stimuli for low-frequency words drawn 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 freely viewed stimuli from conditions 1, 2, 3, and 4 in a counter-balanced design (described below), and each stimulus was presented for 5 seconds. The free viewing task has been widely used in previous studies (e.g., Judd et al, 2009; Cerf et al., 2009). The software "Eyetrack" developed by Jeffrey D. Kinsey, David J. Stracuzzi, and Chuck Clifton, University of Massachusetts Amherst, was used for recording eye movements.

Analysis. Two eye movement measures were taken: *correlation (R)* and *Receiver Operating Characteristic (ROC)*. The Pearson correlation coefficient R between two maps is computed according to sampling points taken every 10 pixels along the x and y axes, and then the correlation coefficient between saliency/center-bias/text-detector and attentional maps (described below) are obtained. An example of a stimulus image and its attention, saliency, center-bias, and text-detector maps is shown in Figure 2. The computation of the ROC measure is described in Hwang, Higgins & Pomplun (2009). If a map had higher correlation or ROC values with regard to the subjects' fixations, the map was considered a better predictor of visual attention. The chance level is 0.5 for ROC and 0 for R.

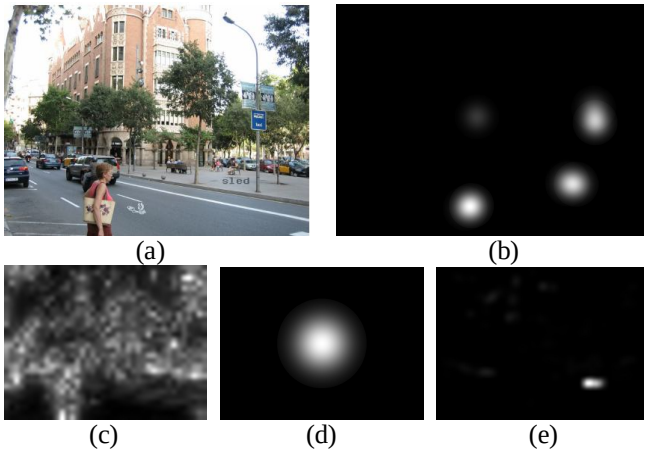


Figure 2. An example of (a) stimulus image, (b) attention (3-second viewing) (c) saliency, (d) center-bias, and (e) text-detector maps.

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. A center-bias map was obtained using a two-dimensional Gaussian distribution at the center of the screen with 3 degrees of visual angle (90 pixels in our experiment setting). The text-detector maps were computed using the automatic text detector which analyzes features such as variation of edge width and edge density.

For the attentional map, we excluded the initial center fixation and included all other fixations within a given viewing duration. The attentional map was built according to each fixation in an image by a two-dimensional Gaussian distribution centered at the fixation point, where the standard deviation was one degree of visual angle to approximate the size of the human fovea. Then we simply summed up these Gaussian distributions for fixations weighted by their durations (see Pomplun, Ritter, & Velichkovsky, 1996).

We computed the attentional maps for each image inspected by each viewer for the initial 1.5, 2, …, 5 seconds. The averages of correlations and ROC values for each viewer were calculated for all, text-present, text-absent, text in front of homogeneous (H-BG), and text in front of inhomogeneous backgrounds (INH-BG) images, and an ANOVA and paired t-tests were performed to analyze the differences between these values

Results and Discussion

Models with and without Text-Detector Maps. The average R and ROC values of all 12 viewers are shown in Table 2. Text-detector maps overlap attentional maps the best when the images contain text in front of homogeneous background, and the worst in text-absent images. These results are consistent with the finding by Judd et al. (2009) that object detectors by themselves do not predict attention well when the objects are absent and therefore should be used in conjunction with other features.

Table 2: The average R and ROC of saliency (Sali), center-bias (Center), text-detector (TextDet), saliency combined with center-bias (SC), and all combined (SCT) maps as predictors of the attentional maps for 3-second viewing. H-BG represents images in front of homogeneous background, and INH-BG represents images on inhomogeneous background.

	Sali	Cen	TextDet	SC	SCT
R -All	0.14	0.16	0.15	0.18	0.20
Text-Present	0.11	0.12	0.20	0.14	0.16
H-BG	0.09	0.10	0.24	0.10	0.12
INH-BG	0.14	0.15	0.15	0.17	0.19
Text-Absent	0.15	0.19	0.12	0.21	0.22
ROC - All	0.65	0.63	0.63	0.69	0.72
Text-Present	0.61	0.61	0.66	0.64	0.70
H-BG	0.55	0.60	0.67	0.58	0.67
INH-BG	0.67	0.62	0.64	0.70	0.72
Text-Absent	0.67	0.64	0.62	0.72	0.73

One-way ANOVAs with the factor "predictor" showed that the performances of Sali, Cen, TextDet, SC, and SCT maps differed significantly in all, text-present, H-BG, INH-BG, and text-absent images for R, all $Fs(4; 55) > 3.64$, $ps < .05$, and ROC, all $Fs(4; 55) > 11.17$, $ps < .01$. SC (without text-detector) obtained significantly lower measures than SCT (with text-detector maps) for all, text-present, H-BG, INH-BG, and text-absent images for R, all $ts(11) > 3.93$, $ps < .01$, and ROC, all $ts(11) > 7.68$, $ps < .001$. The results indicate that the text detector improved the prediction of viewers' visual attention. It is interesting to see that the SCT obtained higher R and ROC than the SC even in text-absent images. One plausible explanation is that some non-objects containing text-like features catch a disproportionate amount of attention.

Text-Present vs. Text-Absent and H-BG vs. INH-BG Images. The five predictors were analyzed in one-way ANOVAs with the factor "image type," and the results

demonstrate that both R and ROC values significantly differed in all, text-present, text-absent, H-BG, and INH-BG images, all Fs(4; 55) > 4.91, ps < .01, and all Fs(4; 55) > 4.72, ps < .01, respectively, except ROC for Cen, F(4; 55) = 0.92, p > .4. The text detector (TextDet) performed better for text-present images than text-absent ones with regard to R, t(11) = 10.67, p < .001 as well as ROC, t(11) = 5.66, p < .001. Homogeneous background images obtained higher values than inhomogeneous background images for both R, t(11) = 7.31, p < .001, and ROC, t(11) = 3.94, p < .01.

Visual Attention over Time. SCT outperformed SC (without text detector) for all viewing durations for R and ROC in both text-present images, both ts(11) > 9.68, ps < .001, and text-absent ones, both ts(11) > 3.93, ps < .01. The difference between SCT and SC was larger in text-present images than in text-absent ones. In text-present images, the R of TextDet initially dominated but decreased over time, while the R of Sali increased (see Figure 3a).. These data suggest that texts are typically detected early during the inspection process and receive sustained attention while the viewers are reading them, thereby elevating the occurrence of text features near fixation. Later in the process, viewers tended to be guided more strongly by saliency as defined by the Itti and Koch algorithm. In text-absent images, the R of Sali, Cen, and TextDet increased over time, indicating that the corresponding mechanisms became more important during the later – likely more focused and fine-grained (Unema, Pannasch, Joos, & Velichkovsky, 2005) – stages of inspection. Clearly, Sali and Cen played more important roles when texts are absent.

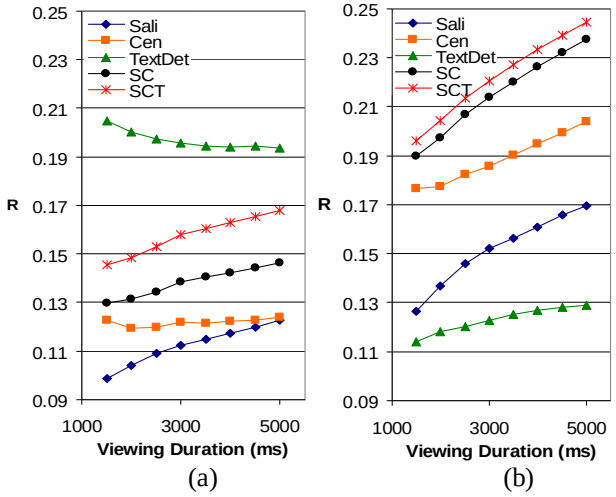


Figure 3. Correlations for 1.5-, 2-, …, and 5-second viewing of (a) text-present and (b) text-absent images.

Experiment 2: English vs. Chinese Texts and Native Speakers

In Experiment 1, we showed that the addition of a text-detector map to saliency and center-bias maps makes the model a better predictor of viewers' visual attention. Our hypothesis is that viewers have developed a "text detector" because they are exposed to texts everyday and become sensitive to text-patterns. Wang and Pomplun (under revision) found that native speakers of English and Chinese-speakers were both attracted by English and Chinese texts in real-world scenes but were attracted more strongly by the texts of their native languages. The reason might be that English and Chinese texts share some common features, such as the histogram of edge width, but also contain their unique features, e.g., Chinese texts usually contain vertical, horizontal, and diagonal strokes but fewer "curves" (such as in "O" or "G" in English). In Experiment 2, the dataset in Wang and Pomplun (submitted) was reanalyzed and our expectation was that the text detector (Lu, submitted) designed for English texts will perform better prediction of gaze fixations for English-speaking viewers than for Chinese-speaking ones.

Method

Participants. In the group of non-Chinese English speakers, 14 students from the University of Massachusetts at Boston participated. All of them were native speakers of English, and none of them had learnt any Chinese or had participated in Experiment 1. 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. At both sites, the experiment setup was identical to Experiment 1.

Stimuli. As shown in Figure 4, the original texts were either rotated by 180 degrees or replaced by Chinese texts. 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. Figure 4a 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 4b, 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 texts.

Figure 4. Example of Chinese and upside-down English texts used in Experiment 2. (a) Condition C1 (b) Condition C2.

Procedure. The procedure was identical to Experiments 1 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.

Analysis. The analyses were identical to Experiment 1.

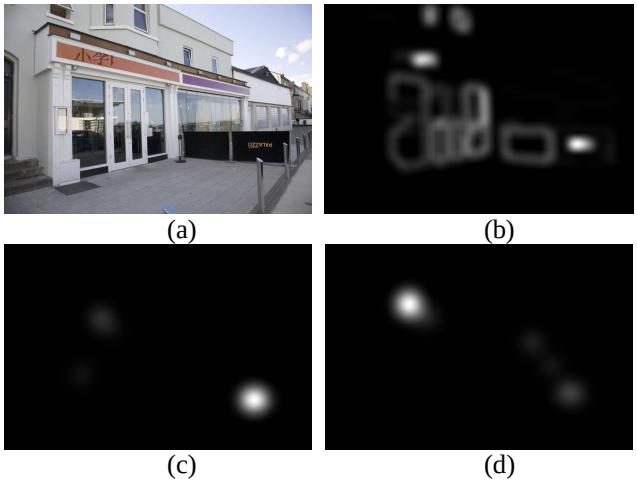


Figure 5. An example of (a) stimulus image, (b) text-detector map, (c), attentional map of an English-speaking viewer (5-second viewing), and (d) attentional map of a Chinese-speaking viewer (5-second viewing).

Results and Discussion

Models with and without Text-Detector Maps. The average R and ROC of all 14 English-speaking and 16 Chinese-speaking viewers are shown in Table 3. For English-speaking viewers, one-way ANOVAs showed that the Sali, Cen, TextDet, SC, and SCT maps performed differently in all, text-present, and text-absent images for R, all $Fs(4; 65) > 8.47$, $ps < .01$, and for ROC, all $Fs(4; 65) > 53.78$, $ps < .001$. SCT predicted attentional maps better than SC in all, text-present, and text-absent images for R, all $ts(13) > 3.49$, $ps < .01$, and ROC, all $ts(13) > 6.61$, $ps < .001$. For Chinese-speaking viewers, similar results were obtained - the performances of Sali, Cen, TextDet, SC, and SCT maps significantly differed for both R, all $Fs(4; 75) > 33.91$, $ps < .001$, and ROC, all $Fs(4; 75) > 22.86$, $ps < .001$. SCT yielded better prediction of attentional maps than SC for both R, all $ts(15) > 4.85$, $ps < .001$, and ROC, all $ts(15) > 5.29$, $ps < .001$. The results of SCT are consistent with Experiment 1 in that the text detector improved the prediction of viewers' visual attention, even in text-absent images.

Text-Present vs. Text-Absent Images. For English-speaking viewers, TextDet performed better in text-present images than in text-absent ones for both R, $t(13) = 6.41$, $p < .001$, and ROC, $t(13) = 5.58$, $p < .001$. For Chinese-speaking viewers, similar results were found: text-present images obtained higher R and ROC than text-absent ones, $t(15) = 4.97$, $p < .001$, and $t(15) = 7.35$, $p < .001$, respectively.

English vs. Chinese-Speaking Viewers. As shown in Figure 6, TextDet predicted English-speaking viewers' attention better than Chinese-speaking viewers' attention for all viewing durations in both text-present images, $t(7) = 23.12$, $p < .001$, and text-absent images, $t(7) = 5.38$, $p < .01$. These results indicate that the text detector that was designed for English texts performed better at predicting the allocation of attention for English-speaking viewers than for Chinese-speaking ones.

Table 3: The average R and ROC of saliency (Sali), center-bias (Cen), text-detector (TextDet), saliency combined with center-bias (SC), and all combined (SCT) maps as predictors of attentional maps for 5-second viewing. En represents English-speaking viewers, and Ch means Chinese-speaking viewers.

	Sali	Cen	TextDet	SC	SCT
R (En)	0.17	0.17	0.14	0.20	0.21
Text-Present	0.15	0.16	0.16	0.19	0.21
Text-Absent	0.18	0.17	0.12	0.21	0.22
R (Ch)	0.17	0.16	0.12	0.19	0.20
Text-Present	0.15	0.15	0.14	0.18	0.19
Text-Absent	0.18	0.17	0.11	0.20	0.21
ROC (En)	0.69	0.61	0.60	0.72	0.73
Text-Present	0.68	0.62	0.63	0.71	0.73
Text-Absent	0.69	0.61	0.59	0.72	0.73
ROC (Ch)	0.68	0.60	0.60	0.70	0.71
Text-Present	0.67	0.61	0.62	0.69	0.71
Text-Absent	0.68	0.60	0.58	0.70	0.70

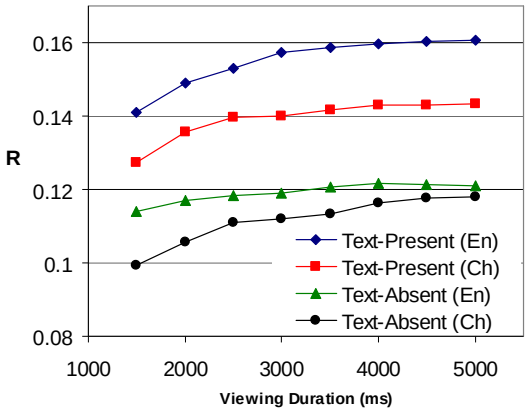


Figure 6. The R values of TextDet for 1.5-, 2-, …, and 5-second viewing of text-present and text-absent images by English-speaking (En) and Chinese-speaking (Ch) viewers.

General Discussion

In Experiment 1, we found that adding a text detector to an attention model improved its prediction of viewers' visual attention, even in text-absent images. Our results suggest that non-text objects whose features resemble those of texts (such as high spatial frequency edges) catch a disproportionate share of attention. Based on the current

data, it seems that the viewers' "biological text detectors" are somewhat similar to the artificial system and influence the viewers' distribution of attention when viewing real-world images. From a time-course analysis, it appears that the biological text detector influences the allocation of attention particularly strongly during later stages of image inspection when viewers are increasingly likely to attend to detailed local structures (see Unema et al., 2005) for semantic interpretation of perceived text.

Whereas the results of Experiment 1 could have been caused by the text detection algorithm being sensitive to visual features that generally attract attention, such as edge density, this interpretation becomes implausible given the results of Experiment 2. We found that the text detector designed for English texts predicted English-speaking viewers' attention better than Chinese-speaking viewers', supporting the hypothesis that viewers have developed a "text detector" that is sensitive to text patterns they are familiar with. It is interesting to see that the way we learn to read influences our allocation of visual attention in everyday life, even when there are no texts presented and we are not specifically looking for any texts.

While the present study has demonstrated the influence of language on visual attention in real-world scenes, further research needs to identify the visual features that underlie this effect. This could be achieved by using text detection algorithms for different writing systems and test their individual components as predictors of native and non-native speakers' attention in natural scenes. Besides a more comprehensive understanding of attentional control in humans, such studies may also result in technological advances. Human viewers can easily locate texts in natural scenes, performing clearly better than current text-detection techniques even when the texts are degraded by noise, rotated, distorted, or shown from unusual perspectives. Consequently, the results of this line of research, such as analyzing what features or local structures are actually learned by the biological text detector, might contribute to the development of more effective automatic text detectors, which could, for example, make a great difference to visually challenged people's lives.

Acknowledgments

Preparation of the article was supported by Grant R01EY021802 from the National Eye Institute to Marc Pomplun.

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.

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.

Epshtein, B., Ofek, E., & Wexler Y. (2010). Detecting text in natural scenes with stroke width transform. *Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 2963–2970.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A Discriminatively Trained, Multiscale, Deformable Part Model. *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, 1-8.

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 (25).

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.

Jung, C., Liu, Q., and Kim, J. (2009). A stroke filter and its application for text localization. *Pattern Recognition Letters*, 30(2):114–122.

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.

Lu, S., Wang, H.-C., J.-H. Lim, & Pomplun, M. (submitted). Learning Text Saliency for Automatic Text Detection in Natural Scenes.

Pomplun, M., Ritter, H., & Velichkovsky B., (1996). Disambiguating Complex Visual Information: Toward Communication of Personal Views of a Scene, *Perception*, 25, 8, 931-948.

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.*

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B.M. (2005). Time course of information processing during scene perception. *Visual Cognition*, 12(3), 473-494.

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

Wang, H. C. & Pomplun M. (2011). The attraction of visual attention to texts in real-world scenes. *The Annual Meeting of the Cognitive Science Society (Cogsci2011)*, 2733–2738.

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, 504-513.