Vanishing point attracts gaze in free-viewing and visual search tasks

Ali Borji
Center for Research in Computer Vision, Department of Computer Science, University of Central Florida, Orlando, USA

Mengyang Feng
Department of Electrical Engineering, Dalian University of Technology, Dalian, China

Huchuan Lu
Department of Electrical Engineering, Dalian University of Technology, Dalian, China

Several structural scene cues such as gist, layout, horizontal line, openness, and depth have been shown to guide scene perception (e.g., Oliva & Torralba, 2001); Ross & Oliva, 2009). Here, to investigate whether vanishing point (VP) plays a significant role in gaze guidance, we ran two experiments. In the first one, we recorded fixations of 10 observers (six male, four female; mean age 22; $SD = 0.84$) freely viewing 532 images, out of which 319 had a VP (shuffled presentation; each image for 4 s). We found that the average number of fixations at a local region (80×80 pixels) centered at the VP is significantly higher than the average fixations at random locations (t test; $n = 319$; $p < 0.001$). To address the confounding factor of saliency, we learned a combined model of bottom-up saliency and VP. The AUC (area under curve) score of our model (0.85; $SD = 0.01$) is significantly higher than the base saliency model (e.g., 0.8 using attention for information maximization (AIM) model by Bruce & Tsotsos, 2005, t test; $p = 3.14\text{e-}16$) and the VP-only model (0.64, t test; $p < 0.001$). In the second experiment, we asked 14 subjects (10 male, four female; mean age 23.07, $SD = 1.26$) to search for a target character (T or L) placed randomly on a 3×3 imaginary grid overlaid on top of an image. Subjects reported their answers by pressing one of the two keys. Stimuli consisted of 270 color images (180 with a single VP, 90 without). The target happened with equal probability inside each cell (15 times L, 15 times T). We found that subjects were significantly faster (and more accurate) when the target appeared inside the cell containing the VP compared to cells without the VP (median across 14 subjects 1.34 s vs. 1.96 s; Wilcoxon rank-sum test; $p = 0.0014$). These findings support the hypothesis that vanishing point, similar to face, text (Cerf, Frady, & Koch, 2009), and gaze direction (Borji, Parks, & Itti, 2014) guides attention in free-viewing and visual search tasks.

Introduction

Visual attention is crucial in understanding complex scenes and processing an enormous amount of information (around 10^8 bits) bombarding our retina each second. Primates use focal visual attention and rapid eye movements to analyze complex visual inputs in real-time, in a manner that highly depends on behavioral priorities and goals. Studies of physiology and psychophysics have proposed that several factors such as *bottom-up cues*, *nature of the target*, and *knowledge of the task* play important roles in guiding attention and eye movements. In what follows, we briefly explain these cues.

Bottom-up cues guide attention based on low-level image-based features. Such cues make a red dot more salient among a set of blue dots. According to the Feature Integration Theory (FIT) by Treisman and Gelade (1980), several feature maps such as color, orientation, and intensity are computed in parallel across the visual field and are then combined to guide the attention. Koch and Ullman (1987) later introduced a computational architecture to generate a master saliency map and proposed a selection process to sequentially deploy attention to spatial locations in decreasing order of their salience. Several computational attention models have been proposed since then to detect bottom-up salient regions that stand out from their surroundings in an image (Borji, Sihite, & Itti, 2013b; Borji & Itti, 2013; Bruce & Tsotsos, 2005; Itti, Koch, & Niebur, 1998). These models have been shown to reliably predict fixations in free viewing of natural scenes (Borji et al., 2013b; Borji, Frintrop, Sihite, & Itti,

Citation: Borji, A., Feng, M., & Lu, H. (2016). Vanishing point attracts gaze in free-viewing and visual search tasks. *Journal of Vision*, *16*(14):18, 1–22, doi:10.1167/16.14.18.

doi: 10.1167/16.14.18 Received December 5, 2015; published November 30, 2016 ISSN 1534-7362

2012; Bruce & Tsotsos, 2005; Parkhurst, Law, & Niebur, 2002).

Prior knowledge of the target facilitates target detection in visual search tasks (See Borji, Lennartz, & Pomplun, 2015; Chen & Zelinsky, 2006; Zelinsky, 2008). The guided search theory (Wolfe, 2007) proposes that attention can be biased towards targets of interest by modulating the relative gains through which different features contribute to attention. Psychophysics experiments have shown that knowledge of the target contributes to an amplification of its salience. For example, Blaser, Sperling, and Lu (1999) report that white vertical lines become more salient if we are looking for them. Some studies have shown that better knowledge of the target leads to faster search; e.g., seeing an exact picture of the target is better than seeing a picture of the same semantic type or category as the target (Kenner & Wolfe, 2003; Maxfield, Stalder, & Zelinsky, 2014). Physiology experiments have shown that target search modulates neural activity by enhancing the responses of neurons tuned to the location and features of a stimulus (Bichot, Rossi, & Desimone, 2005; Martinez-Trujillo & Treue, 2004; Saenz, Buracas, & Boynton, 2002; Treue & Trujillo, 1999). Navalpakkam and Itti have investigated computational and behavioral underpinnings of these processes (2005, 2006, 2007) and have proposed models to bias the low-level visual system with the known features of the target to make the target more salient (also known as the optimal gain theory). Borji and Itti (2014b) have studied how parameters of neurons in a neural population should be optimally biased to search for a target.

Further, it has been shown that attention in the real world is mainly task-driven (Ballard, Hayhoe, & Pelz, 1995; Borji & Itti, 2014a; Hayhoe & Ballard, 2005). The classic eye movement experiments of Yarbus, Haigh, and Riggs (1967) show drastically different patterns of eye movements over the same scene, depending on the task. He demonstrated a striking example of how a verbally communicated task specification may dramatically affect attentional deployment and eye movements. He argued that variable spatio-temporal characteristics of scanpath for different task specifications exemplify the extent to which behavioral goals may affect eye movements and scene analysis. Another example in this regard is the study by Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995) who investigated the interplay between task demands (spoken sentence comprehension) and gaze control by tracking eye movements of subjects when they received ambiguous verbal instructions regarding manipulating objects on a table. Tanenhaus et al. demonstrated that visual context influenced spoken word recognition and syntactic processing when subjects had to resolve ambiguous verbal instructions by analyzing the visual scene and objects. These two studies indicate that visual attention and scene understanding are intimately interrelated and that gaze is controlled by task demands. In another work, Triesch, Ballard, Hayhoe, and Sullivan (2003) suggested that our brain may adopt a need-based approach, where only desired objects are quickly detected in the scene, identified, and represented. In natural vision, bottom-up saliency, search template (object features), scene context and layout, and task demands interact with each other in guiding visual attention.

The geometry of a scene provides global contextual information that assists rapid scene analysis in visual search and navigation (Ross & Oliva, 2009). Structural cues such as layout, depth, openness, and perspective can be perceived in a short presentation of an image (Greene & Oliva, 2009; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Sanocki & Sulman, 2009; Schyns & Oliva, 1994). Further, global scene context or Gist[1] (Torralba, Oliva, Castelhano, & Henderson, 2006) and layout[2] (Rensink, 2000) also guide attention to likely target locations in a top-down manner. Given a task such as "find humans in the scene," in addition to visual features representing the appearance of a person, the gist of the scene also guides the search process. For example, humans are more likely to be found on the sand, rather than on the sky, in a beach scene. Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009) proposed a model to linearly integrate three components (bottom-up saliency, gist, and object features) for explaining eye movements in looking for people in a database of about 900 natural scenes. Structural scene information has also been extensively modeled and utilized in several computer vision applications, for example, geometry context (Hoiem, Efros, & Hebert, 2005) and Manhattan world (Coughlan & Yuille, 2003). Whereas influences of structural information on visual recognition (e.g., rapid scene categorization and understanding) have been studied, their role in visual attention has not yet been systematically explored.

In the context of driving, previous research has shown that drivers rely on the road tangent point (the point of the inner lane marking bearing the highest curvature in the 2D retinal image) when negotiating a bend. The bend radius (and hence the steering angle) relates in a simple geometrical fashion to the visible angle between the instantaneous heading direction of the car and the tangent point (Land & Lee, 1994). Drivers can easily use this strategy by looking at the tangent point and inferring the required steering angle from the rotation angle of their gaze and head. Driving by the tangent point has been observed from both normal and racing drivers in real world scenarios (Chattington, Wilson, Ashford, & Marple-Horvat, 2007; Land & Lee, 1994; Land & Tatler, 2001), as well as in simulated conditions (Wilson, Chattington, &

Marple-Horvat, 2008). Another study has found that drivers look straight ahead at the road 59% of the time, to the right side of the road 15% of the time, and to the left side of the road 25% of the time (Ko, Higgins, Chrysler, & Lord, 2010). Underwood, Chapman, Brocklehurst, Underwood, and Crundall (2003) showed that the far-ahead road and the midahead road attract more fixations than other parts of the scene. Further, they found that road near-left and the road near-right tend to attract fewer fixations.

In summary, several cognitive cues that attract attention and guide eye movements in natural vision have been already discovered (e.g., color, texture, motion, face, and text (Cerf et al., 2009), object center-bias (Nuthmann & Henderson, 2010), scene center-bias (Tatler, Baddeley, & Gilchrist, 2005), cultural cues (Chua, Boland, & Nisbett, 2005; Rayner, Castelhano, & Yang, 2009), semantic distance (Hwang, Wang, & Pomplun, 2011), and gaze direction (Borji et al., 2014; Castelhano, Wieth, & Henderson, 2007). Further, structural scene information such as global context, horizontal line,[3] and openness (Torralba et al., 2006), scene layout (Rensink, 2000), and depth (Le Meur, 2011; Ouerhani & Hügli, 2000) have been shown to influence eye movements as well as human scene categorization (Friedman, 1979; Potter, 1976; Oliva & Torralba, 2001). Here, we systematically investigate the role of a particular type of scene structural information known as the vanishing point (VP) and perspective on eye movements during free-viewing and visual search in natural scenes. In a graphical perspective, a *vanishing point* is a 2D point (in the image plane) which is the intersection of parallel lines in the 3D world (but not parallel to the image plane). In other words, the vanishing point is the spot to which the receding parallel lines diminish. In principle, there can be more than one vanishing point in the image. VP can commonly be seen in fields, railroads, streets, tunnels, forest, buildings, objects such as ladder (from looking bottom-up), etc. It has been used in camera calibration, 3D reconstruction as well as in painting.

Experiment 1: Free viewing

Methods

Stimuli

We collected 700 images with vanishing points from the Web, MIT300 (Judd, Durand, & Torralba, 2012) and DUTOMRON (Yang, Zhang, Lu, Ruan, & Yang, 2013) datasets. We used a vanishing point detector based on the Hough transform (Ballard, 1981) to discard images with multiple VPs or VPs falling off the image plane. The VP detector did well in localizing the VP for some images and failed for some others. We also discarded low resolution images and images with logos. Eventually, we were left with 319 images each with only one vanishing point. Overall, we attempted to create a diverse stimulus set containing images from roads, buildings, beaches, corridors, tunnels, forest, indoor, outdoor, natural, man-made, etc., with vanishing points in multiple locations.

We asked two subjects to annotate the VP location.[4] They were shown images with the maximum resolution[5] of 400×300 pixels, and were asked to mark the VP with a bounding box of an arbitrary size. The average height and width of the VP bounding boxes were 10 and 14 pixels, respectively. In the latter analyses, we only use the center of the bounding boxes.

Since showing only images with a VP may guide observers to adopt a viewing strategy, we collected additional 213 images without vanishing points and shuffled them among images with VPs. Therefore, viewers would not know in advance whether a presented image would contain a VP. In total, we had 532 images to record human fixations (319 with VP and 213 without). In what follows, only images containing vanishing points are analyzed. For presenting to subjects, images were resized to 1920×1080 pixels by adding gray margins to them while preserving the aspect ratio.

Figure 1A shows examples of our stimuli, annotated vanishing points, as well as fixation locations. Figure 1B shows the average annotation map as well as the average fixation map over 319 images with VPs. Both of these maps have maximum activation near the image center making center-bias a potential confounding factor for our hypothesis. We will address this confound extensively in our analyses. Figure 1C shows the histogram of VP window size (bounding box). About 82% of annotated VP bounding boxes have a size smaller than 0.2% the image area.

Observers

Observers were undergraduates from different majors (six male, four female). Mean observer age was 22 (range 21–24, median = 22, $SD = 0.84$). Observers had normal or corrected-to-normal vision and received course credits for participation. They were naive to the purpose of the experiment and had not previously seen the stimuli.

Procedure

An image was shown for 4 s followed by a gray screen for 3 s. Observers sat 60 cm away from a 19″ LCD monitor such that scenes subtended approximately $37.6° \times 24°$ of the visual field. A chin rest was used to stabilize head movements. Stimuli were

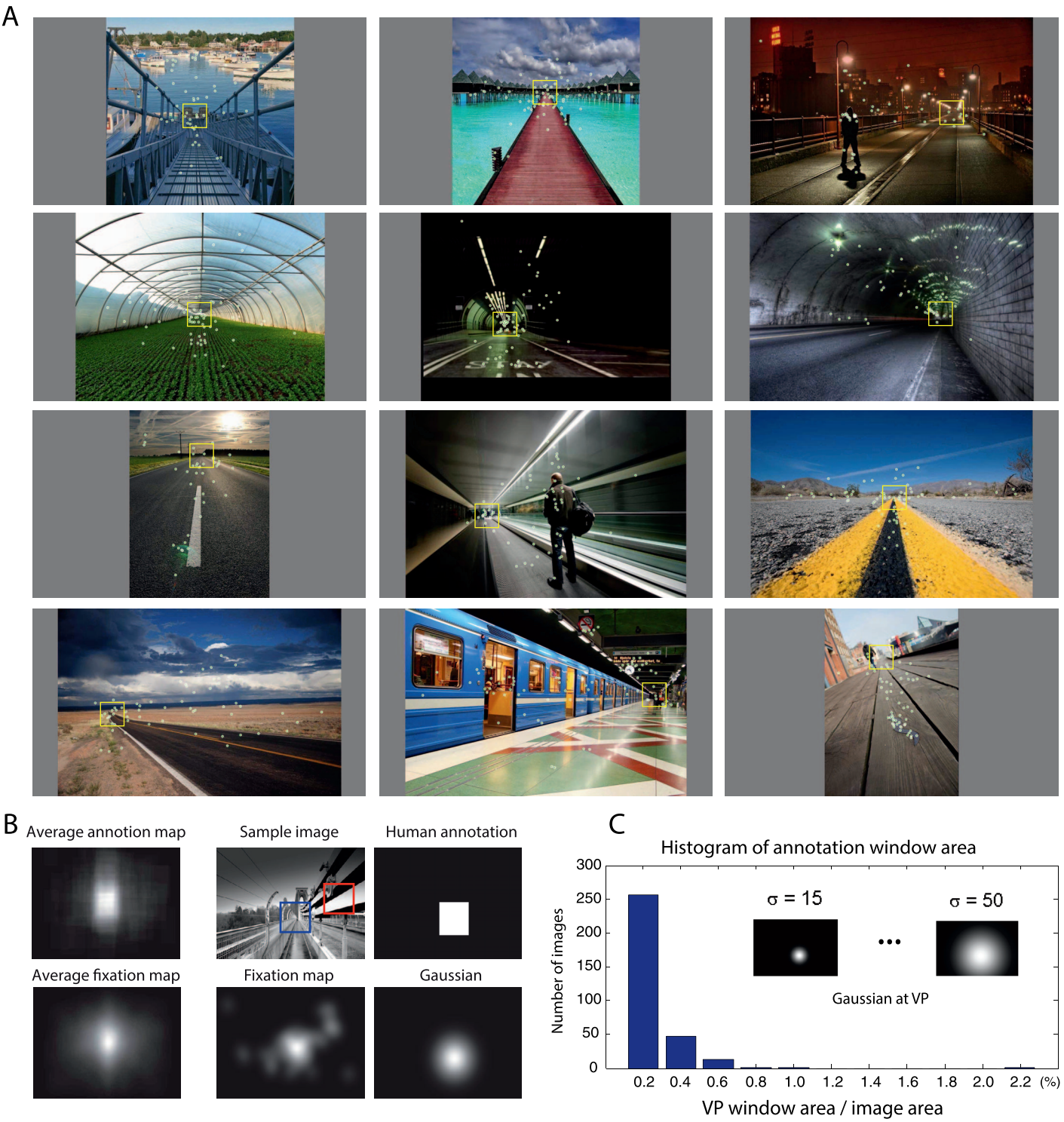


Figure 1. (A) Example stimuli with vanishing points (yellow boxes) and fixations (dots) used in Experiment 1. These images were shown to observers shuffled among some other images without vanishing point to avoid viewing bias (i.e., strategy in viewing). For images with some highly salient items, vanishing point attracts less attention (e.g., second image in the fourth row). (B) Average VP annotation map and average fixation map over 319 images with VPs. A sample image and its corresponding human VP annotation, eye movements of 10 observers, and Gaussian blob centered at the VP is also shown. The inset shows two squares with size 80×80 pixels (blue = square centered at the VP location, red = square centered at the VP location of another randomly chosen image). (C) Histogram of VP window size. About 82% of VP bounding boxes have a size smaller than 0.2% of the image area. Inset shows two Gaussian blobs with smallest and largest σ values.

presented at 60 Hz at a resolution of 1920 × 1080 pixels (with added gray margins while preserving the aspect ratio). Eye movements were recorded via a Tobii X1 Light Eye Tracker at a sample rate of 30 Hz. The eye tracker was calibrated using five-point calibration at the beginning of each recording session. Images were presented to observers in a random order. Observers were instructed to simply watch and enjoy the pictures (free viewing task).

Model-free analysis

In our first analysis in this section, we compute and compare the density of fixations inside the VP and random bounding boxes of the same size (80 × 80 pixels). We take the center of the annotated rectangles and draw an 80 × 80 square at that location.[6] To generate a random bounding box for an image, we use the VP bounding box of another randomly chosen image. This way, random locations have the same central bias as the VPs.

The average number of fixations inside the VP squares is 33.1 ($SD = 15.5$) which is significantly higher than the average number of fixations at random locations (19.8; $SD = 16.2$) using the t test ($n = 319$; p = 2.98 e-35). This implies that fixations are driven to vanishing points (see Figures 2A, 2B).

In our second analysis, we repeat the first analysis with saliency maps (Figure 2). We find that average saliency, using Itti (Itti et al., 1998), AIM (Bruce & Tsotsos, 2005), and Boolean map saliency (BMS) (Zhang & Sclaroff, 2013) models in the VP square, is significantly higher than random squares (t test; $n = 319$; p = 4.23e-13 using the Itti model, $p = 2.03$e-08 using the AIM model, and p = 0.02 using the BMS model). Please see Figures 2C and 2D. One may argue that the higher number of fixations and map activations using the human fixation map at vanishing points compared to random points could be because of higher saliency (as computed by conventional bottom-up saliency models) around vanishing points (as shown above). In other words, feature congestion at VP induces higher low-level image-based saliency and drives attention. This makes bottom-up saliency a potential confounding factor for our hypothesis here. To address this confound, we measured the ratio of average activation at VP versus random locations using saliency models as well as the blurred human fixation map (by convolving with a small Gaussian blob of $\sigma =$ 10 pixels; Figure 1B). The ratio using the human fixation map is 1.5, which is higher than ratios using models (1.3 for Itti, 1.2 for AIM, 1.1 for BMS). This means that while saliency plays a role, it cannot fully account for the VP effect. We will investigate this further in the next section by utilizing a model-based analysis.

To investigate how early is the bias towards the vanishing point, we calculate the ratio of first fixations that landed on the VP square over the total number of first fixations (from all subjects on an image) and compare it with the same ratio using second fixations (these ratios were computed for each image and were then averaged over all images). The former value is 0.45 which is significantly higher than the latter value of 0.38 (t test; $p < 0.001$). This indicates that observers were driven to the vanishing point early in their viewing. Figure 3 shows the heatmaps of first and second fixations on some example images.

Model-based analysis

In the previous analyses, we showed that saliency at the VP is higher than random locations. To investigate whether attraction of fixations towards vanishing point is solely due to the higher saliency at VP or there is a significant additional value, we perform a model-based analysis. If the vanishing point offers redundant information to what saliency already offers, then explicit emphasis on the vanishing point should not add significant prediction power. Similar analyses have been pursued in the past to study whether other types of cues such as face or text (Cerf et al., 2009), object center-bias (Borji, 2015; Borji & Tanner, 2015), and gaze direction (Borji et al., 2014; Parks, Borji, & Itti, 2015) guide eye movements and attention.

Learning a combined model of saliency and vanishing point

To learn a combined model of saliency and vanishing point, we represent each image pixel as a vector $\mathbf{X} = [s\ v]$ where s is the output of a bottom-up saliency model[7] and v is the value from the vanishing point map (VP). The VP map is modeled as a variable size Gaussian placed at the vanishing point as shown in Figure 1.[8]

$$VP(x, y) = \frac{1}{2\pi\sigma_{vp}^2} e^{-\frac{(x-i)^2+(y-j)^2}{4\sigma_{vp}^2}} \quad (1)$$

where (i, j) is the coordinate of the annotated vanishing point and σ_{vp} is the (variable) standard deviation of the Gaussian blob.

We aim to learn function $\phi(\mathbf{X}) = W^T\mathbf{X} + b$, which is a binary classifier determining whether a particular image pixel with feature vector $\mathbf{X}$ should be attended or not—i.e., $\phi(\mathbf{X}) \in \{-1, 1\}$. To do so, we utilize an ordinary support vector machine (SVM; Cortes & Vapnik, 1995)

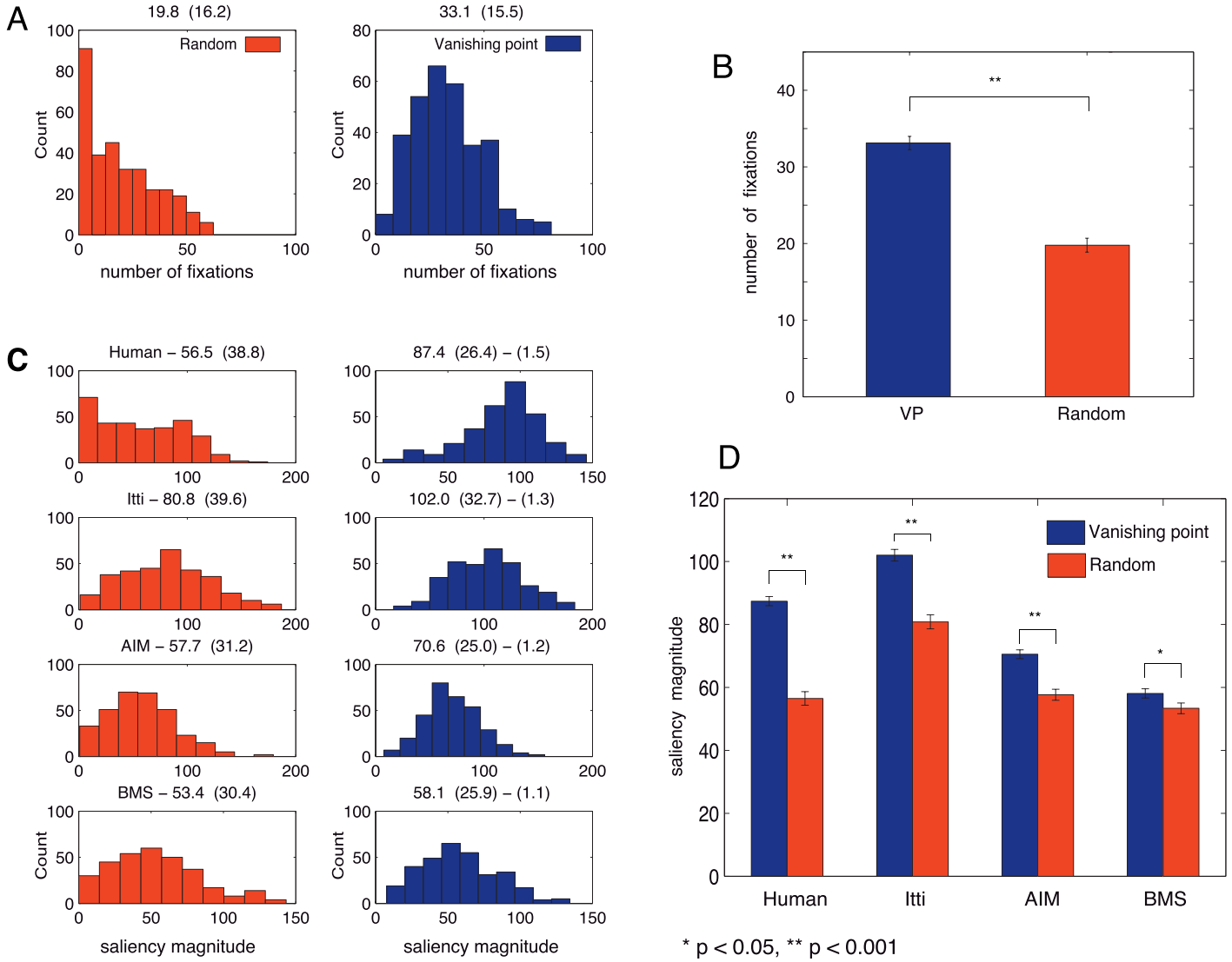


Figure 2. Results of the first experiment. (A) Histogram of fixations at random (left column) and annotated VP squares (80 × 80 pixels). Plot titles show mean and standard deviations (in parentheses). (B) The mean number of fixations at VP and random squares (sampled with the same bias as VP annotations to account for center-bias). The difference is statistically significant using *t* test ($n = 319$). (C) Histogram of saliency map activations at VP and random squares (see Figure 1B) using human saliency map (fixation map convolved with a small Gaussian blob with $\sigma = 10$ pixels). The second parenthesis on the title of the right bar column shows the ratio of the means (i.e., mean-at-VP/mean-at-random). (D) Differences in VP and random squares means. All differences are statistically significant using *t* test. Error bars show standard error of the mean *(SEM)*.

with a linear kernel. For a test pixel, we assign the real value $m = W^T\mathbf{X} + b$ as the label of the pixel.[9] Final saliency values are then normalized for each map, that is, (m − min)/(max − min). We deliberately avoid using complicated nonlinear learning functions, since we are interested in the exclusive added value of the vanishing point.

We choose 50 random images to train the SVM and use the remaining 269 images for testing. This procedure is repeated 20 times and then the average is computed (i.e., cross-validation). We randomly select 50 pixels respectively from fixated locations and non-fixated locations, yielding 100 samples (50 positive samples and 50 negative samples) for each training image, i.e., 5,000 training samples in total. We learn the combined models (e.g., AIM + VP, BMS + VP, and Itti + VP) and compare them with the original bottom-up saliency models. For a fair comparison, we optimize models by sweeping[10] σ_{vp} from 15 to 50 pixels and find the σ where performance peaks.[11] Please see Figure 4A.

Model evaluation

Table 1 summarizes the test results using three types of scores: area under the curve (AUC), normalized

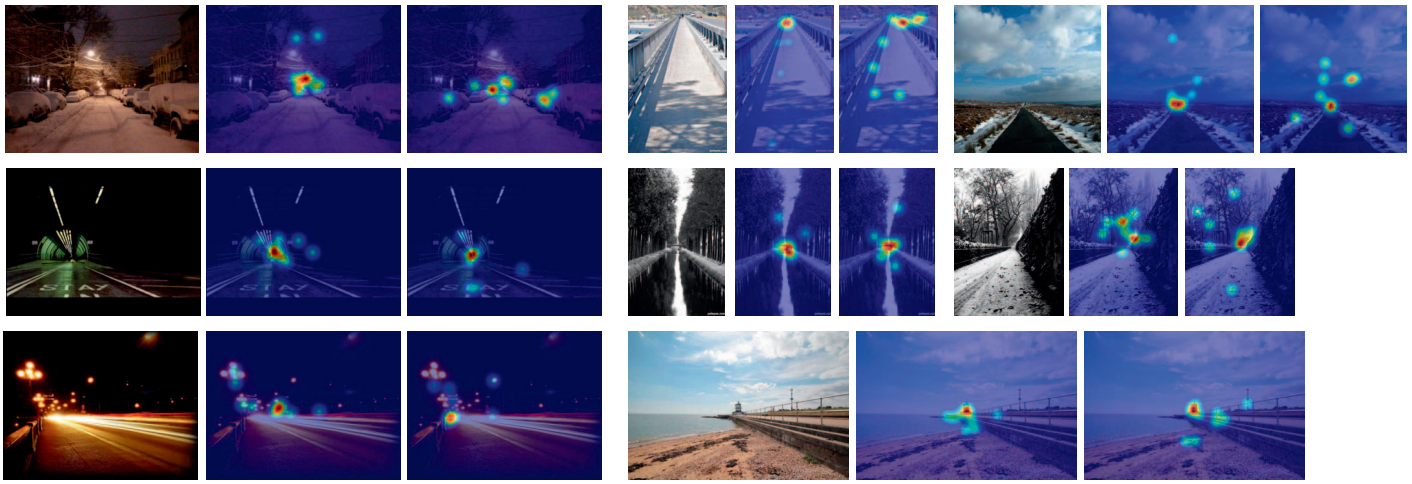


Figure 3. Sample images with distribution of first and second fixations.

scanpath saliency (NSS) proposed by Peters, Iyer, Itti, and Koch (2005), and linear correlation coefficient (CC). Please see Borji and Itti (2013) for a detailed definition of these scores in the context of fixation prediction. We make three observations explained below (see Figure 4A).

First, we observe that the VP performs significantly above chance in predicting fixations using all three scores (Table 1, last column). Using the AUC score, VP map offers at least 43% improvement versus chance.

Second, adding VP to models significantly outperforms baseline models using three scores (M +VP vs. M; Table 1, fifth column). We observe more than 8.5% improvement in performance using the AUC score (more than 48% using NSS and more than 52.5% using CC). This result confirms our hypothesis that VP offers significant additional prediction power than bottom-up saliency.

Third, the M + VP model outperforms the VP model using the AUC score but performs slightly lower than the VP using NSS and CC scores due to a different amount of activation generated by these models (Table 1, fourth column). This comparison is not very relevant to our hypothesis here as it compares VP and bottom-up saliency. Marginal improvement of M + VP over VP (and sometimes lower performance), hints towards the strong attraction of gaze towards the vanishing point such that emphasizing more on bottom-up salient items degrades the accuracy. Nonetheless, results from fourth and fifth columns of Table 1 confirm that both bottom-up saliency and vanishing point contribute statistically significantly in guiding attention and gaze in free viewing and none in a subset of the other one.

Addressing center-bias

It has been shown that human observers tend to preferentially look near the center of the image due to reasons such as viewing strategy or photographer bias (the tendency of photographers to frame interesting objects at the center of the picture). This phenomenon is known as center bias and has challenged researchers in testing hypotheses in eye movement studies and comparing saliency models (Borji et al., 2013a; Tatler et al., 2005; Tseng, Carmi, Cameron, Munoz, & Itti, 2009). In this section, we address this confounding factor in detail. Notice that VP happens at the center of some of our images.

We perform two comparisons. First, we investigate whether VP and CG (Central Gaussian) models offer nonredundant information. Notice that both models have been optimized for their best σ. The scatter plot in Figure 4B shows that for 97 of 269 test images, the VP model wins over the CG model. Inspecting images in which VP wins, we see that VP falls off the image center and observers look at the VP location. For images where CG wins, there is either: (a) a salient object at the image center, (b) a VP at the image center, or (c) the scene contains mainly background clutter. In the latter case, observers tend to look at the center as there is not much interesting in the periphery to look at.

Second, we compare the M + CG + VP model versus the M + CG model (parameters optimized for each model separately). The rationale behind this comparison is to see whether VP boosts the performance of the augmented (with CG) model. Figure 4C shows performance increase for 221 images (about 82% of images). The third column in Table 1 shows the quantitative results and statistical tests (t test; $n = 269$). The difference between M + CG + VP and M + CG models is statistically significant using all bottom-up saliency models and scores. These two comparisons confirm that vanishing point and central bias are two different nonoverlapping cues that attract fixations and guide attention.

We also compute another score known as the shuffled AUC (sAUC) score (Zhang, Tong, Marks,

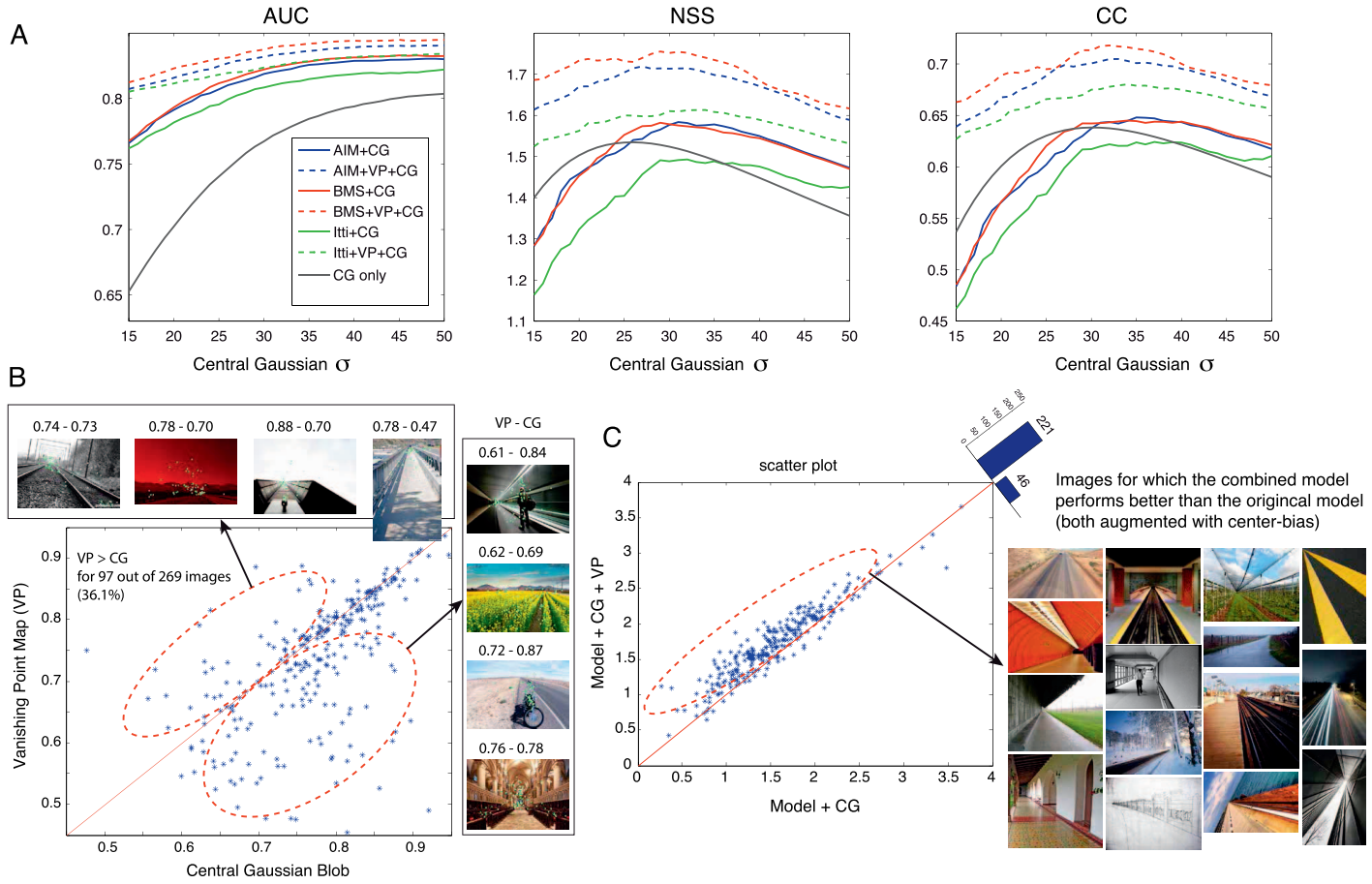


Figure 4. (A) The performance of models using AUC, NSS, and CC scores with varying Gaussian σ. The x axis indicates the σ of the central Gaussian blob. The σ of the VP Gaussian blob is fixed to the best sigma for each model. (B) AUC score of the VP map versus the CG central Gaussian (CG) map using the best σ (here both σ values are 31 pixels). Each dot represents one image. Some examples above and below diagonal are shown. (C) Scatter plot of Model + CG versus Model + CG + VP using the NSS score. This plot shows the added value of VP over the original model, taking into account the center bias confound. This plot also indicates that the added value is not due to the center-bias. For 221 images, we observed performance improvement. Vanishing point usually happens off the center on these images. We did the same analysis by plotting the Model + VP versus M and observed performance improvement for 243 of the images (not shown).

Score	Model = M	M + CG + VP versus M + CG	M + VP versus VP	M + VP versus M	VP versus Chance
AUC	AIM	0.833 versus 0.819 (1.7%)	0.793 versus 0.739 (7.3%)	0.793 versus 0.720 (10.13%)	0.739 versus 0.5 (47.8%)
	BMS	0.837 versus 0.823 (1.7%)	0.798 versus 0.719 (11%)	0.798 versus 0.711 (12.2%)	0.719 versus 0.5 (43.8%)
	Itti	0.826 versus 0.811 (1.84%)	0.792 versus 0.756 (4.7%)	0.792 versus 0.730 (8.5%)	0.756 versus 0.5 (51.2%)
NSS	AIM	1.695 versus 1.545 (9.7%)	1.467 versus 1.450 (1.2%)	1.467 versus 0.953 (53.9%)	1.450 versus 0
	BMS	1.751 versus 1.587 (10.33%)	1.543 versus 1.508 (2.3%)	1.543 versus 0.916 (68.4%)	1.508 versus 0
	Itti	1.592 versus 1.445 (10.2%)	1.361 versus 1.381 (−1.4%)	1.361 versus 0.918 (48.25%)	1.381 versus 0
CC	AIM	0.697 versus 0.628 (11%)	0.584 versus 0.598 (−2.3%)	0.584 versus 0.358 (63.12%)	0.598 versus 0
	BMS	0.720 versus 0.652 (10.42%)	0.609 versus 0.608 (0.16%)	0.609 versus 0.341 (78.5%)	0.608 versus 0
	Itti	0.672 versus 0.603 (11.44%)	0.563 versus 0.580 (−2.9%)	0.563 versus 0.369 (52.6%)	0.580 versus 0

Table 1. Scores of our combined model (Model + VP) versus the original model and the VP-only channel. *Notes*: Numbers in parentheses are the performance improvement in percentages. Scores of the center-bias augmented models are also shown (third column). Differences are all statistically significant (t test; $n = 269$; $p < 0.05$) except the CC score of M + VP versus VP using the Itti model ($p = 0.83$).

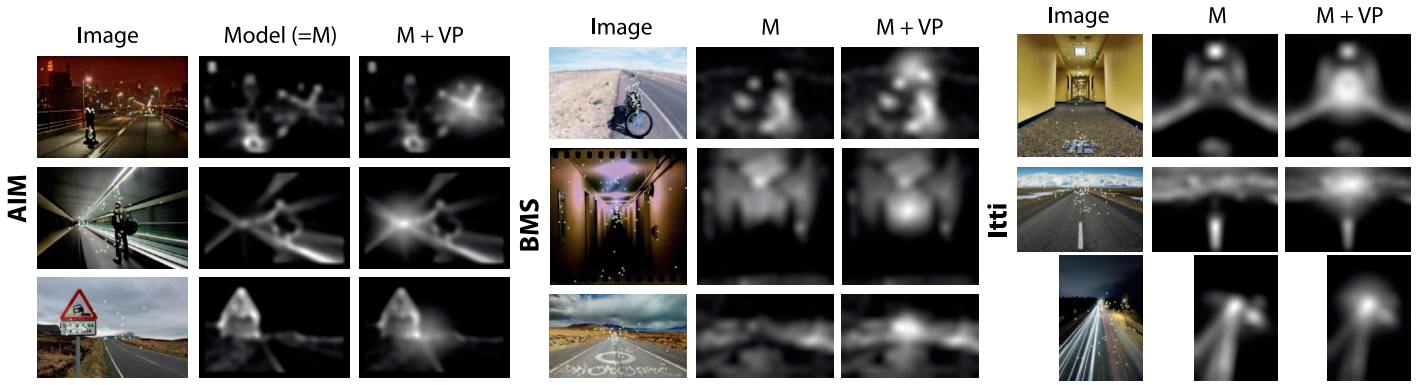


Figure 5. Qualitative evaluation: cases where our combined model performs poorly. In almost all of these cases, an object off the vanishing point overrides the VP effect and attracts fixations.

Shan, & Cottrell, 2008) to compare M + VP versus M (Model). The only difference between this score and the classic AUC score is that negative samples, instead of uniform locations, are drawn from fixations on other images to discount the center bias (see Borji et al., 2013a, for a detailed explanation of these scores). We learn a weighted linear model first, and then optimize its VP sigma over 50 random train images. The learned model is then tested over the remaining 269 test images and the process is repeated five times. The average sAUC score of the combined model is significantly above the base saliency models (t test; $p < 0.05$; scores in order using AIM, BMS, and Itti models are 0.658 vs. 0.626, 0.671 vs. 0.631, and 0.644 vs. 0.61). Note that there is no center-bias Gaussian addition here. See Appendix B for more details.

Qualitative evaluation

Here, we evaluate model prediction maps qualitatively. Figure 5 shows examples where our combined model performs poorly (i.e., M + VP scores lower than M). In almost all of these cases, an object off the vanishing point overrides the VP effect. For example, in the second row of the first column, a standing person in the metro station attracts gaze more than VP (the same is true for images in first row, second column and first row, third column).

Figure 6 shows examples where our combined model performs well. Scores of models are also shown. The original baseline models (AIM, BMS, and Itti) fail to render the vanishing point salient in almost all of the shown cases. Augmented with VP, however, we see an improvement in the prediction power (Compare M + VP vs. M). M + CG model also outperforms the M model. The best performance is achieved using the M + VP + CG model. This model however sometimes loses to the M + VP model as adding CG causes false positives in some cases. Comparing the VP and CG models (the last two columns) indicates that VP sometimes wins over CG, especially in cases where VP is off the image center.

Complementary analysis

In a complementary analysis, to investigate whether vanishing point guides attention in other tasks, we evaluate the performance of our combined model over existing datasets in the literature. Fourteen images with vanishing points were selected from the FIGRIM (Bylinskii, Isola, Bainbridge, Torralba, & Oliva, 2015) dataset and VP locations were annotated. FIGRIM images have been used for memorability testing to probe whether eye movements relate to whether a subject will later remember an observed image. Note that these images have been shown to observers among many other images without VP. Therefore, there is no bias towards looking at the vanishing point on these datasets. The performance of our combined model over these images is shown in Table 2. As it can be seen, our model performs better than original models using all three scores. Figure 7 shows an illustration over the FIGRIM dataset.

Experiment 2: Visual search

Our aim in this experiment is to investigate whether our findings from the free-viewing task generalize to other tasks. In particular, we attempt to see whether presence of a vanishing point attracts attention during visual search.

Methods

Stimuli

Our stimuli contain 270 color images, selected from images used in Experiment 1, 180 of which with vanishing

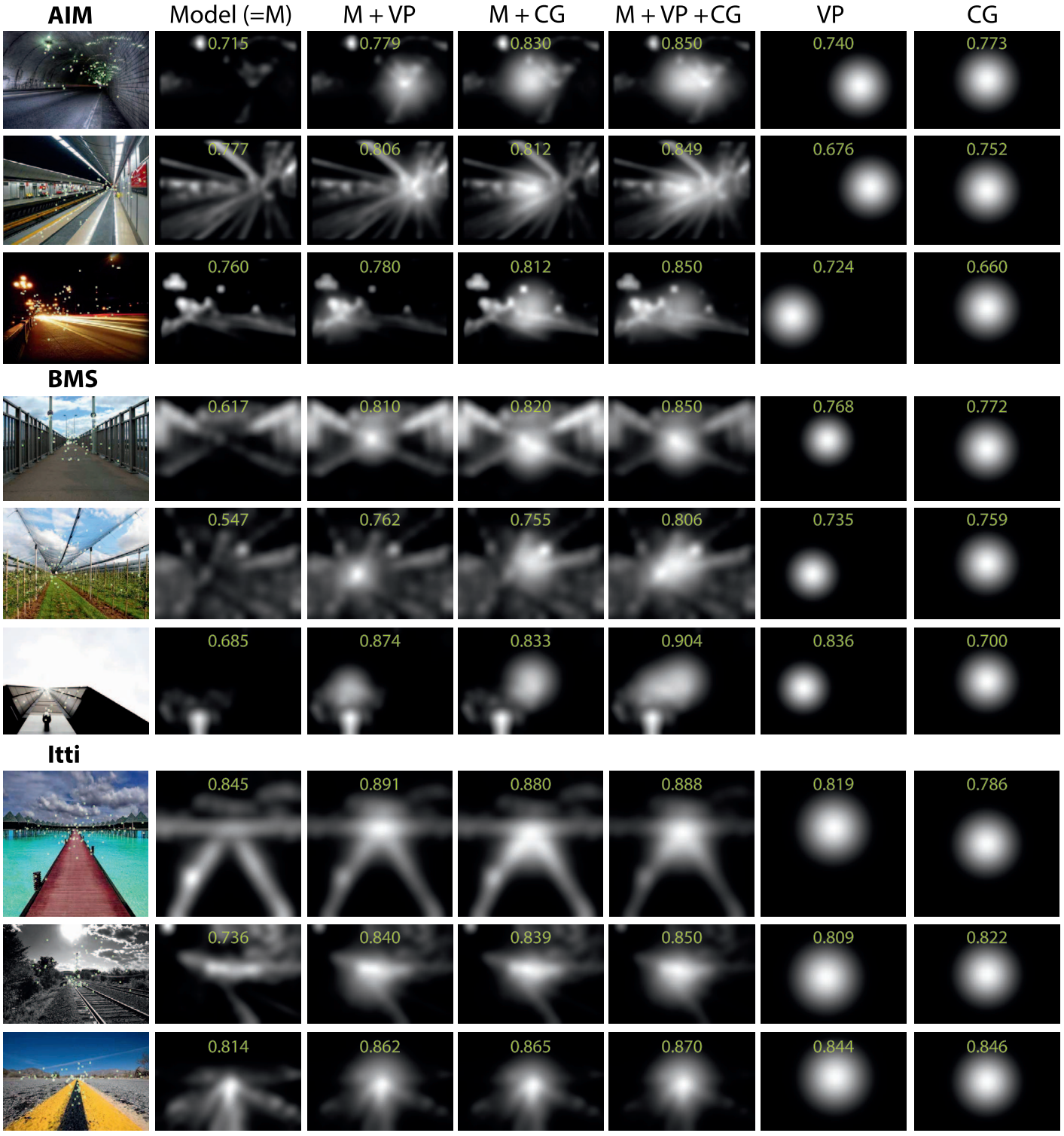


Figure 6. Qualitative evaluation: cases where our combined model performs well (using the AUC score).

points and 90 without. Images without VP were used to reduce the viewing bias towards the vanishing points. A target character (T or L) was placed randomly on a 3×3 imaginary grid overlaid on an image (see Figure 8B, fourth row, first column). The target character occupies $0.80° \times 0.67°$ of the visual field and is placed at the center of a grid cell. The target happened with equal probability (1/9) inside each cell (30 times: 15 times L, 15 times T). Twenty times, out of this 30 happened on images with a vanishing point (10 L, 10 T) and 10 times over images without a VP (5 L, 5T). Figure 8 shows example stimuli used in the visual search experiment.

Score model	AIM + VP	BMS + VP	ITTI + VP	SALICON + VP
AUC	0.837 (0.758)	0.854 (0.764)	0.846 (0.799)	0.825 (0.795)
NSS	1.731 (1.081)	1.851 (1.151)	1.716 (1.221)	2.611 (2.379)
CC	0.511 (0.308)	0.533 (0.321)	0.523 (0.363)	0.683 (0.614)
Best VP σ	31	27	37	16
W	[9.453, 6.693]	[10.716, 6.982]	[6.883, 6.931]	[16.222, 6.164]

Table 2. Maximum scores of our combined models over 14 images with vanishing points selected from the FIGRIM (Bylinskii et al., 2015) dataset. *Notes*: Numbers in parentheses are scores of the original models.

Subjects

Fourteen subjects (10 male, four female) voluntarily participated in this experiment. Mean subject age was 23 (range 22–27, median = 23, $SD = 1.27$). All subjects were undergraduates from different majors. They were naive to the purpose of the experiment and had not previously seen the stimuli.

Procedure

As in Experiment 1, observers sat 60 cm away from a 19″ LCD screen. Stimuli were presented at 60 Hz at a resolution of 1920 × 1080 pixels with added gray margins (dva 37.6° × 24°). Subjects were asked to search for the target character. They were instructed to report their answers by pressing one of two arrow keys (left arrow for T, right arrow for L). If no key was pressed after 10 s, the trial automatically moved to the next one. Subjects were not asked to look at the center of the screen before each trial. The reason was to avoid preference towards the center where vanishing point usually happens.[12] Each stimulus was succeeded by a gray screen for 4 s. Overall, it took about 30 min for a subject to complete this task. We measured subjects' response times and accuracies.

Analysis and results

Figure 9 shows the results of our second experiment. Analyzing images with VP, we found that all subjects[13] were significantly faster when the target happened inside the cell containing the VP compared to off-VP cells (median across 14 subjects, 1.34 s vs. 1.96 s; Wilcoxon rank-sum test; $p = 0.0014$; Figure 9A). Median (and also mean) accuracies over subjects were above 95% (Figure 9B). Accuracies for the target in VP cells were significantly higher than off-VP cells (medians 100% vs. 97%; $p = 0.02$). This result supports our hypothesis that vanishing point guides attention during visual search.

To investigate whether center-bias is a confounding factor here (since VP often happens near the image center), we conducted two analyses. First, we computed the mean distance of the target cell (center of the cell containing the target) from the image center in VP and off-VP trials. These two values in order are: 442.12 (120.9) and 428.36 (193.54). Numbers in parentheses are standard deviations. As can be seen, on average off-VP cells are closer to the image center (since a larger number of these cells happen near the center compared to cells containing VP). Although off-VP cells are slightly closer to the image center, their response time is still higher than the VP cells. Second, we measured the response time for trials when the target happened at the central cell. The median response time for those trials (across subjects) is 1.94 (mean equal 2) which is higher than response time over VP cells. These results indicate that central bias cannot explain the lower response time towards the vanishing point.

The images without vanishing points were merely used as fillers to reduce the viewing bias. They do not

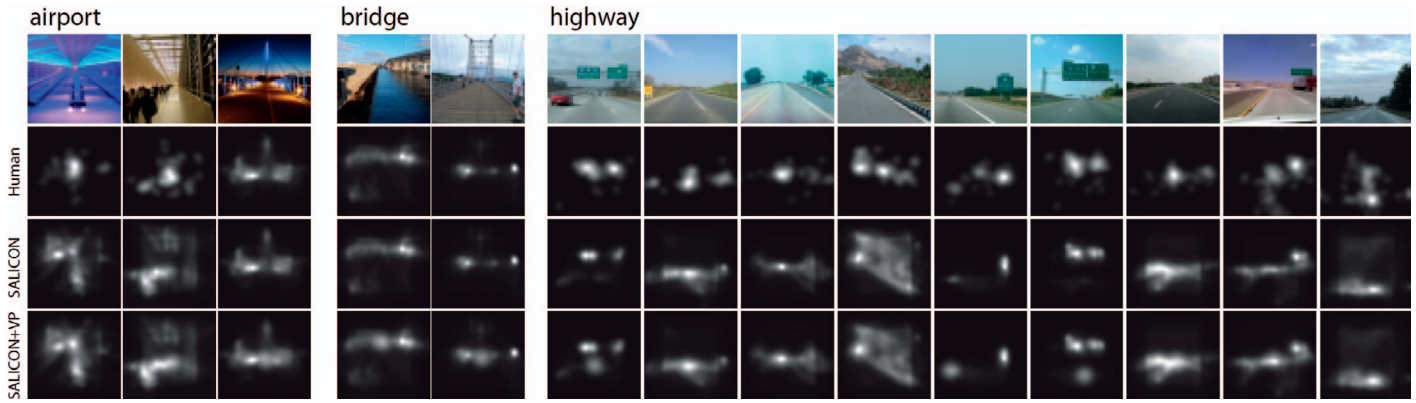


Figure 7. Illustration of predictions of the SALICON model (as one of the best existing saliency models; Bylinskii et al., 2014) versus SALICON + VP model over images taken from the FIGRIM dataset.

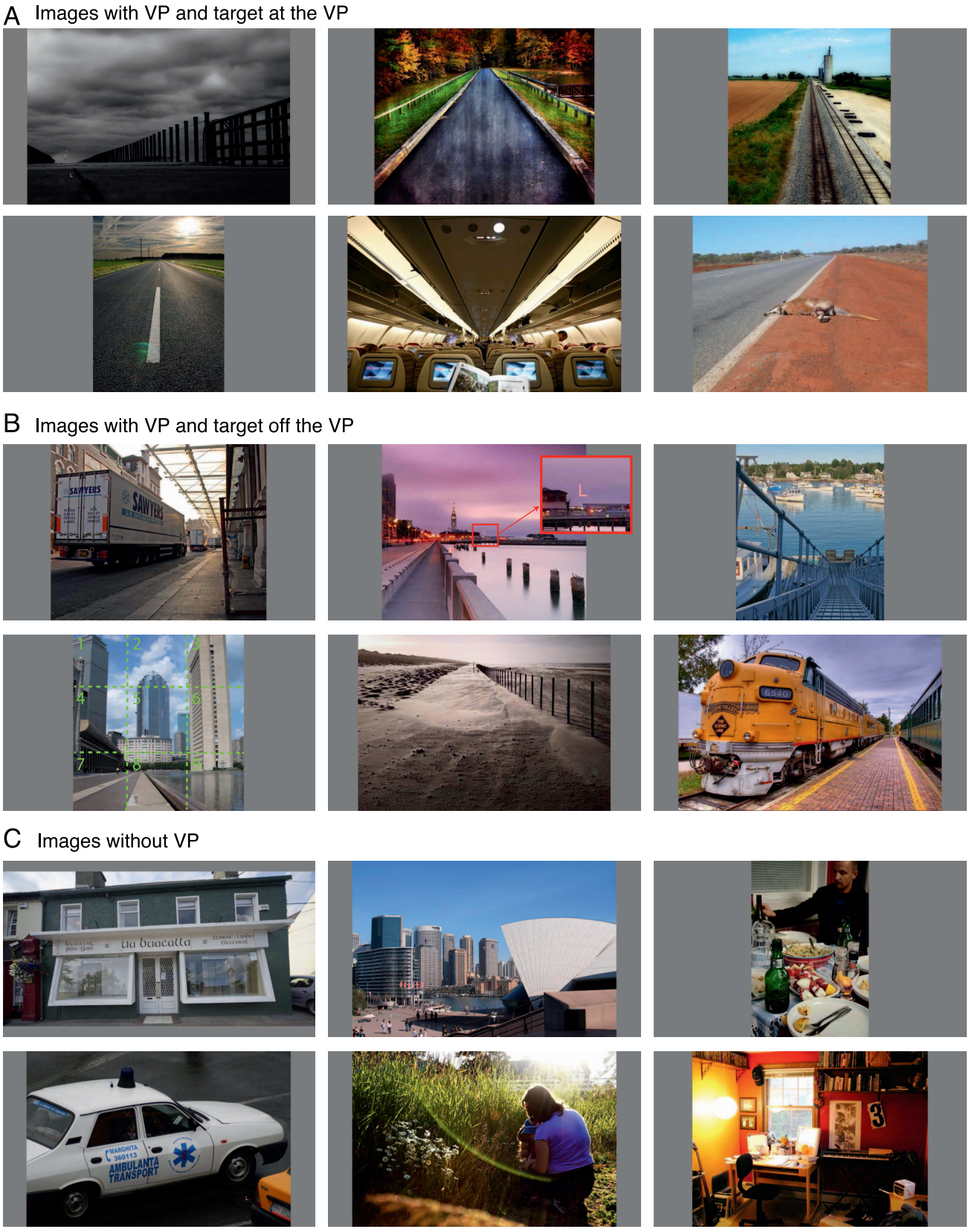


Figure 8. Stimuli used in the visual search experiment (270 images in total). (A) Images with the target at the vanishing point, (B) Images with vanishing point but target off the VP. The image in the fourth row, first column shows a 3 × 3 imaginary grid with the target at the center cell, (C) Images without VP. The inset in the image in the third row, second column, shows the zoomed out target region for better illustration. Try to see if you can locate the grid cell containing the target character (T or L) in images. Answer key: (A) 7, 2, 2, 2, 8, 2; (B) 9, 5, 2, 6, 4, 9; and (C) 4, 5, 7, 9, 7, 7.

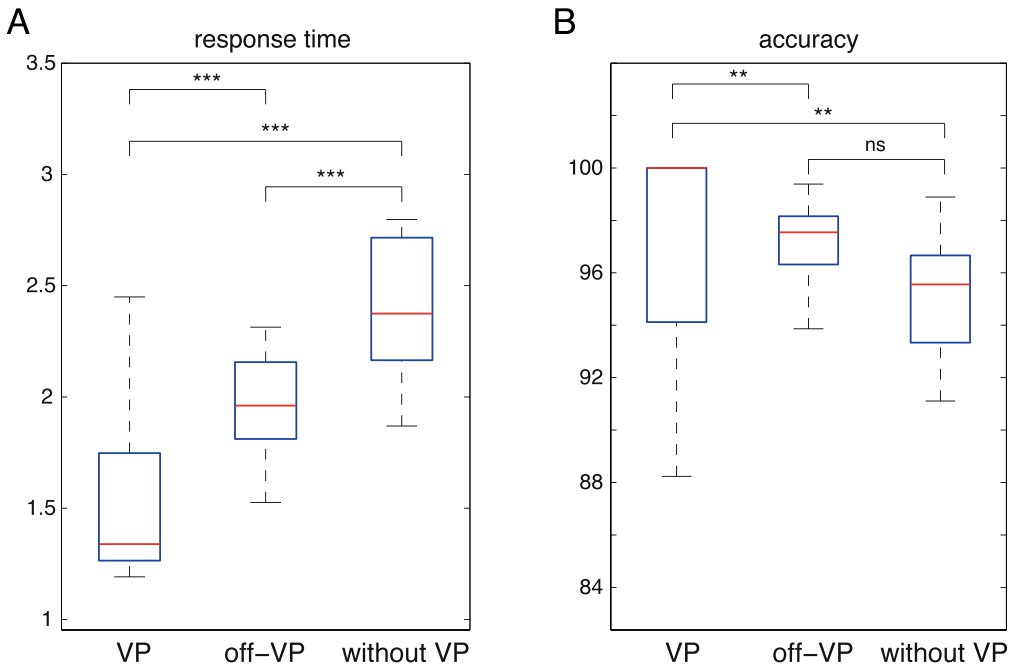


Figure 9. Results of the visual search experiment. (A) Response time. (B) Accuracy. $^{***} = p < 0.001$, $^{**} = p < 0.01$, *ns* = nonsignificant (here $p = 0.055$; almost significant).

influence our results reported above. Nonetheless, we analyzed the data over these images as well. Response time for the images without VP was significantly higher than response time for those with VP (median 2.37; p = 4.77e-05) and off-VP cells ($p = 0.0072$). This does not necessarily mean that, in general, response time on images without vanishing point is higher than images with vanishing point as our finding could be solely due to the statistics of chosen images without VP. Indeed, a further inspection of these images revealed that the majority of them contain faces, text, and other salient stimuli leading to higher complexity and thus higher response time over these images (see, for example, Figure 8C). Notice, however, that the important straight comparison pertaining to our hypothesis is the response time and accuracy in VP versus off-VP cells. The choice of images without vanishing point does not interfere with this comparison.

Discussion

Our data show that there is a viewing guidance towards the vanishing point on images with a strong vanishing point. There might be a continuum between these types of examples and those where the geometry defines a vanishing point that is out of the frame, or wherein the images are less "tunnel-like" (see Figure 10A). It would also be interesting to measure the vanishing point guidance on images with multiple vanishing points (Figure 10B).

A finer-grained analysis of fixations towards the vanishing point might reveal interesting patterns. For example, we notice that subjects have a higher tendency to look beneath the vanishing point (see Figure 11) maybe because of the higher feature density at those locations. In some cases, fixations seem to be marked by a trail that leads towards the vanishing point (e.g., fourth, sixth, seventh, ninth, and 12th images, counting row-wise, in Figure 1) which raises the question of whether an anisotropic profile for this bias factor might produce even stronger results. Emphasizing more on these regions may further improve the prediction power of our combined model.

As we showed in Section 4, explicit addition of a vanishing point channel to a saliency model significantly improves performance on images with vanishing points. Even the best existing saliency model (SALICON by Jiang, Huang, Duan, & Zhao, 2015) according to the MIT saliency benchmark (Bylinskii et al., 2014), trained on a large amount of data, falls short in explaining fixations driven by the vanishing point (see also Bylinskii et al., 2016). The main reason could be because this model has been trained on explicit saliency judgments in which observers were asked to click on salient regions. It is not clear whether subjects are able to discover high-level factors that are known to direct gaze (e.g., gaze direction; Borji et al., 2014). Further, this model has not been trained on many images containing vanishing points. Thus, we believe mining behavioral factors that guide attention and gaze is very valuable in constructing more predictive fixation prediction models.

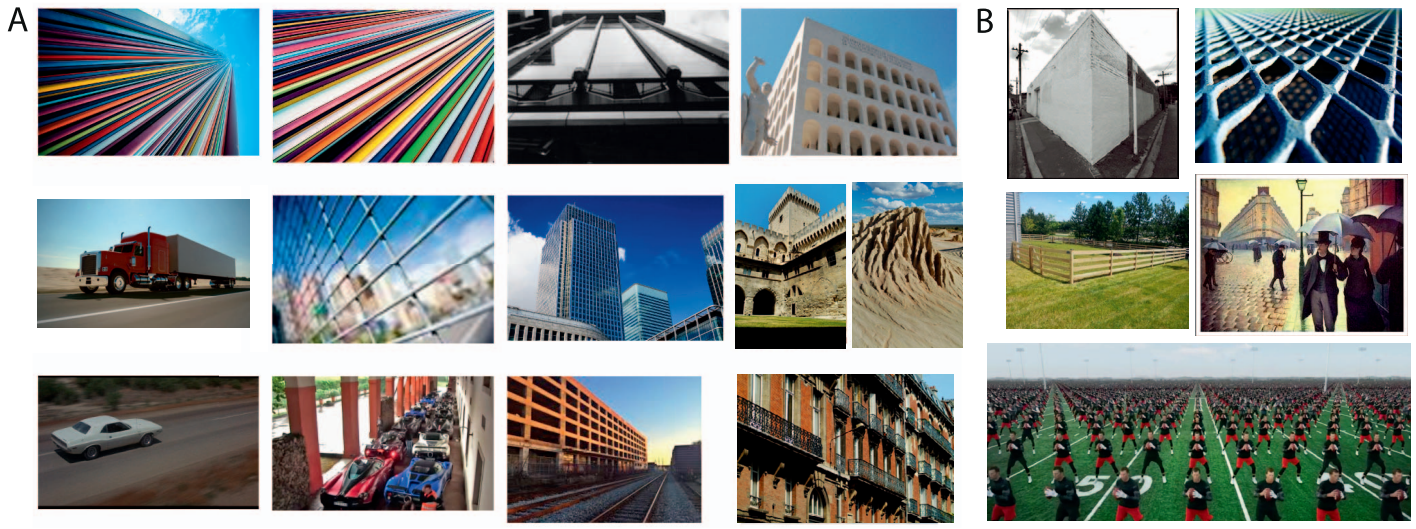


Figure 10. (A) Some images with varying degrees of vanishing point inside the image. Note that in some cases, attention can be attracted to VP areas even though the VP itself falls off the image (e.g., third image in the first row). In some other cases, VP is less likely to capture attention (e.g., fourth image in the third row). (B) Some example images with multiple vanishing points.

Is it a preference to fixate points having a greater perceived distance, or does the VP fixation preference reflect the contribution of a lower level visual factor? By definition, the vanishing point is the point in infinity in the scene (3D world). So, it might be difficult to rule out the depth cue. Note, however, that several points with depth at infinity might exist in the image which may capture less attention compared to the vanishing point (see images in Figure 1). Further, for different sets of parallel lines, their respective vanishing points will lie on a line in the image called *vanishing line* or *horizon line*. This line is obtained by the intersection of the image plane with a plane parallel to the ground plane, passing through the camera center. All vanishing lines end at the horizon line. A likely explanation for the tendency of humans to look near the vanishing point could be a special form of mid- or high-level processing dedicated to vanishing point detection (similar to faces and text). Low-level features such as orientation, color, or intensity alone might not suffice for VP bias. This is one reason why existing bottom-up saliency models fail to account for the vanishing point effect.

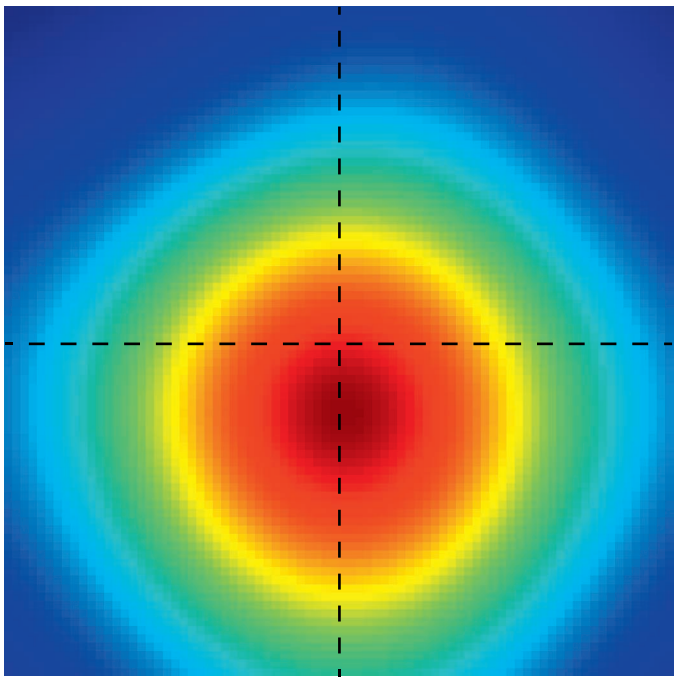


Figure 11. Average fixation map at the vanishing point location (averaged over 319 crops of size 80 × 80 in Experiment 1; see Figure 1). Each of the horizontal and vertical lines halves the image. Subjects tend to look more below the vanishing point.

During the course of this work, two other studies have emerged investigating the role of vanishing point in scene viewing. Deng, Yang, Li, and Yan (2016) addressed gaze in the context of driving. They analyzed the eye-tracking data of 40 nondrivers and experienced drivers when viewing 100 traffic images (road scenes only). They found that a driver's attention was mostly concentrated on the end of the road in front of the vehicle, and bottom-up saliency models can only describe a small portion of the fixations driven by vanishing points. Replicating our prior study (Borji, Feng, & Lu, 2015), Ueda, Kamakura, and Saiki (2016) found that vanishing point attracts fixations during scene free viewing. They also embedded a Gabor patch in a natural scene and asked participants to search for it. Their results show that the first saccade in each trial tended towards a vanishing point. They also found that the vanishing point attracts attention even in scenes

composed of simple geometric figures. These two studies align with our findings in this study.

To explore possible neural mechanisms of vanishing point representation in the visual system, we conducted a synthetic computational experiment. A convolutional neural network (CNN; here Krizhevsky, Sutskever, & Hinton, 2012) was trained to predict whether an image contains a vanishing point or not. After training, we observed that some VP-selective neurons emerged in higher layers of the network. Please see Appendix D for details. A similar observation has been reported in Le (2013) where authors found neurons selective to cat face by training a neural network on unlabeled images. Two findings suggest that, as in CNNs, some VP-selective neurons might also exist in the brain. First, it has been shown that CNNs resemble computational processes underlying object recognition in the visual ventral stream (e.g., Yamins & DiCarlo, 2016). Second, it is known that there are face-selective neurons in the fusiform gyrus (McCarthy, Puce, Gore, & Allison, 1997). Whether such VP-selective cells indeed exist, however, needs to be investigated by careful electrophysiology experiments. Candidate brain regions include parahippocampal place area (PPA), the lateral occipital complex (LOC), as well as other regions in the ventral stream. It has been shown that these regions are involved in the analysis of objects, spatial layout of scenes, and scene geometry (e.g., Park, Brady, Greene, & Oliva, 2011). Altogether, these findings support the hypothesis that devoting some specialized neurons or regions to specific features or categories might be a general design principle in the brain for representation and recognition of complex scenes and objects (see, for example, Leibo, Mutch, & Poggio, 2011).

Conclusion

Previous research has shown that humans are capable of automatic and rapid analysis of scene structure when navigating an environment or searching for objects. It has also been shown that several global scene properties such as *coarse spatial layout* (Schyns & Oliva, 1994), *naturalness* (Joubert et al., 2007), *navigability* (Greene & Oliva, 2009), *complexity or clutter* (Rosenholtz, Li, & Nakano, 2007; Sanocki & Sulman, 2009), *distance and depth* (Sanocki, 2003), and *openness* (Torralba et al., 2006) can be perceived in a short presentation of a scene. Inspired by these findings, we showed that a particular type of scene structure related to the scene layout, known as the vanishing point, strongly influences eye movements in free viewing of natural scenes as well as in visual search. Our results align with the findings that structural scene information influence gaze guidance during visual search (e.g., Henderson, Chanceaux, & Smith 2009) and free-viewing (e.g., Le Meur, 2011) and generalize the previous finding that gaze is guided to the road tangent point during driving (Land & Lee, 1994; Land & Tatler, 2001).

In the first experiment, we showed that the density of fixations around the vanishing point is significantly higher than the density of fixations around random locations. This indicates that observers are more likely to look at objects near the vanishing point. We also proposed a combined model of bottom-up saliency and vanishing point and showed that it outperforms original models. This signifies that vanishing point offers significant additional value than what bottom-up saliency models already offer. Further, we showed that VP performs significantly above chance and cannot be explained by center bias.

In the second experiment, we showed that vanishing point guides attention during visual search and complements other factors involved in target search including spatial context and local object information. Subjects were faster and more accurate when the target character happened near the vanishing point compared to other locations in the image.

Results of our two experiments, together, support the hypothesis that vanishing point, similar to face and text (Cerf et al., 2009) and gaze direction (Borji et al., 2014; Parks et al., 2015) attracts eye movements and attention in free-viewing and visual search tasks and should be considered in constructing more predictive saliency models.

One interesting future research direction is studying neurophysiological underpinnings of vanishing point detection and guidance in the brain using cell recording and fMRI techniques. Exploring the role of other structural scene information in gaze guidance is another direction. Finally, it would be rewarding to find out which cues, among several cues such as face, text, gaze direction, vanishing point, etc., human observers prioritize in paying attention.

Keywords: visual attention, eye movements, bottom-up attention, top-down attention, saliency, free viewing, visual search, vanishing point, perspective, global context, gist, scene perception

Acknowledgments

Commercial relationships: none.
Corresponding author: Ali Borji.
Email: aborji@crcv.ucf.edu.
Address: Center for Research in Computer Vision, Department of Computer Science, University of Central Florida, Orlando, USA.

Footnotes

[1] An abstract spatial representation which is rich enough to recognize the semantic category of a scene, such as indoor office, outdoor beach, street, etc.

[2] A scene can be partitioned into coherent spatial regions based on semantic or visual similarity. For example, a typical beach scene can be represented by three regions: sky on top, water in the middle, and sand at the bottom.

[3] The line that separates the earth from the sky (skyline). Observers are more likely to attend to visual items along the horizontal line (Le Meur, 2011).

[4] Subjects were highly consistent in their selection of vanishing point locations ($R^2 = 0.99$). Please see Appendix A

[5] The largest dimension was resized to 400 pixels. The smaller dimension was resized such as to preserve the aspect ratio and thus may sometimes exceed 300 pixels.

[6] Our investigation with other bounding box sizes results in the same conclusions.

[7] We chose Itti, AIM, and BMS models since they use purely bottom-up cues such as orientation or color and exclude high-level features such as face or text.

[8] We experimentally verified that the Gaussian form of VP works better than square or circle.

[9] This model is essentially similar to learning a linear model: $\alpha\ S + (1 - \alpha)$ VP.

[10] Previous research has shown than smoothing impacts saliency model performance (Borji et al., 2013a; Borji & Itti, 2012).

[11] Please note that for different models, different VP σ leads to the best M + VP performance.

[12] To further alleviate center-bias, we cropped the images in such a way as to distribute the vanishing point over the entire 3×3 grid on the image.

[13] See Appendix C for results over individual subjects.

References

Ballard, D. H. (1981). Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, *13*, 111–122.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks., *Journal of Cognitive Neuroscience*, *7*, 66–80.

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area v4. *Science*, *308*, 529–534.

Blaser, E., Sperling, G., & Lu, Z.-L. (1999). Measuring the amplification of attention. *Proceedings of the National Academy of Sciences, USA, 96*, 11681–11686.

Borji, A. (2015). What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing*, *24*, 742–756.

Borji, A. (2016). Vanishing point detection with convolutional neural networks. *arXiv*, preprint, arXiv:1609.00967.

Borji, A., Feng, M., & Lu, H. (2015). Vanishing point attracts eye movements in scene free-viewing. *arXiv*, preprint, arXiv:1505.03578.

Borji, A., Frintrop, S., Sihite, D. N., & Itti, L. (2012). Adaptive object tracking by learning background context. In *Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE* (pp. 23–30).

Borji, A., & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 478–485).

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*, 185–207.

Borji, A., & Itti, L. (2014a). Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, *14*(3):29, 1–22, doi:10.1167/14.3.29. [PubMed] [Article]

Borji, A., & Itti, L. (2014b). Optimal attentional modulation of a neural population. *Frontiers in Computational Neuroscience*, *8*, 34.

Borji, A., Lennartz, A., & Pomplun, M. (2015). What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. *Neurocomputing*, *149*, 788–799.

Borji, A., Parks, D., & Itti, L. (2014). Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of Vision*, *14*(13):3, 1–32, doi:10.1167/14.13.3. [PubMed] [Article]

Borji, A., Sihite, D. N., & Itti, L. (2013a). Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of Vision*, *13*(10):18, 1–4, doi:10.1167/13.10.18. [PubMed] [Article]

Borji, A., Sihite, D. N., & Itti, L. (2013b). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing, 22*, 55–69.

Borji, A., & Tanner, J. (2015). Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(6), 1214–1226.

Bruce, N., & Tsotsos, J. (2005). Saliency based on information maximization. In *Advances in neural information processing systems*, (pp. 155–162).

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*, 165–178.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2014). MIT saliency benchmark. Available at saliency.mit.edu

Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., & Durand, F. (2016). Where should saliency models look next? In *Proceedings of the European Conference in Computer Vision (ECCV)* (pp. 809–824). Berlin: Springer.

Castelhano, M. S., Wieth, M., & Henderson, J. M. (2007). I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In *Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint* (pp. 251–262). Springer.

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, doi:10.1167/9.12.10. [PubMed] [Article]

Chattington, M., Wilson, M., Ashford, D., & Marple-Horvat, D. (2007). Eye–steering coordination in natural driving. *Experimental Brain Research*, *180*, 1–14.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, *46*, 4118–4133.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences, USA, 102*, 12629–12633.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Coughlan, J. M., & Yuille, A. L. (2003). Manhattan world: Orientation and outlier detection by bayesian inference. *Neural Computation*, *15*, 1063–1088.

Deng, T., Yang, K., Li, Y., & Yan, H. (2016). Where does the driver look? Top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, *17*(7), 2051–2062.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*, 945–978.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*, 316.

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

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*, 188–194.

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, *9*(1):32, 1–8, doi:10.1167/9.1.32. [PubMed] [Article]

Hoiem, D., Efros, A. A., & Hebert, M. (2005). Geometric context from a single image. In *10th IEEE International Conference on Computer Vision, Vol. 1* (pp. 654–661). IEEE.

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

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

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015* (pp. 1072–1080). IEEE.

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

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. CSAIL Technical Reports, TR-2012-001, MIT-CSAIL, 2012.

Kenner, N., & Wolfe, J. M. (2003). An exact picture of your target guides visual search better than any other representation. *Journal of Vision*, *3*(9): 230, doi:10.1167/3.9.230. [Abstract]

Ko, M., Higgins, L., Chrysler, S. T., & Lord, D. (2010). Effect of driving environment on drivers' eye movements: Re-analyzing previously collected eye-tracker data. In *Transportation Research Board 89th Annual Meeting*, 10–1363.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of intelligence* (pp. 115–141). Berlin: Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional

neural networks. In *Advances in neural information processing systems*, (pp. 1097–1105).

Land, M. F., & Lee, D. N. (1994, June 30). Where do we look when we steer. *Nature*, *369*, 742–744.

Land, M. F., & Tatler, B. W. (2001). Steering with the head: The visual strategy of a racing driver. *Current Biology*, *11*, 1215–1220.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8595–8598). IEEE.

Le Meur, O. (2011). Predicting saliency using two contextual priors: The dominant depth and the horizon line. In *2011 IEEE International Conference on Multimedia and Expo* (pp. 1–6). IEEE.

Leibo, J. Z., Mutch, J., & Poggio, T. (2011). Why the brain separates face recognition from object recognition. In *Advances in Neural Information Processing Systems* (pp. 711–719).

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, *14*, 744–751.

Maxfield, J. T., Stalder, W. D., & Zelinsky, G. J. (2014). Effects of target typicality on categorical search. *Journal of Vision*, *14*(12):1, 1–11, doi:10.1167/14.12.1. [PubMed] [Article]

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, *9*, 605–610.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205–231.

Navalpakkam, V., & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 2049–2056). IEEE.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*, 605–617.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8):20, 1–19, doi:10.1167/10.8.20.

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

Ouerhani, N., & Hügli, H. (2000). Computing visual attention from scene depth. In *Proceedings. 15th International Conference on Pattern Recognition, 2000* (Vol. 1, pp. 375–378). IEEE.

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *The Journal of Neuroscience*, *31*, 1333–1340.

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

Parks, D., Borji, A., & Itti, L. (2015). Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. *Vision Research*, *116*, 113–126.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*, 2397–2416.

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

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, and Cognition*, *35*, 254.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*, 17–42.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2):17, 1–22, doi:10.1167/7.2.17. [PubMed] [Article]

Ross, M. G., & Oliva, A. (2009). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1):2, 1–25, doi:10.1167/10.1.2. [PubMed] [Article]

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*, 631–632.

Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology*, *47*, 43–86.

Sanocki, T., & Sulman, N. (2009). Priming of simple and complex scene layout: Rapid function from the intermediate level. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 735.

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

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

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. S., & 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.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*, 575–579.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, *3*(1):9, 86–94, doi:10.1167/3.1.9. [PubMed] [Article]

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7):4, 1–16, doi:10.1167/9.7.4. [PubMed] [Article]

Ueda, Y., Kamakura, Y., & Saiki, J. (2016). *Vanishing points attract eye movements during visual search. Journal of Vision*, *16*(12): 1168, doi:10.1167/16.12.1168. [Abstract]

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., & Crundall, D. (2003). Visual attention while driving: sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, *46*, 629–646.

Wilson, M., Chattington, M., & Marple-Horvat, D. E. (2008). Eye movements drive steering: Reduced eye movement distribution impairs steering and driving performance. *Journal of Motor Behavior*, *40*, 190–202.

Wolfe, J. M. (2007). Guided search 4.0. *Integrated models of cognitive systems* (pp. 99–119). Berlin: Springer.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365.

Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3166–3173). IEEE.

Yarbus, A. L., Haigh, B., & Riggs, A. L. (1967). *Eye movements and vision* (Vol. 2). New York: Springer.

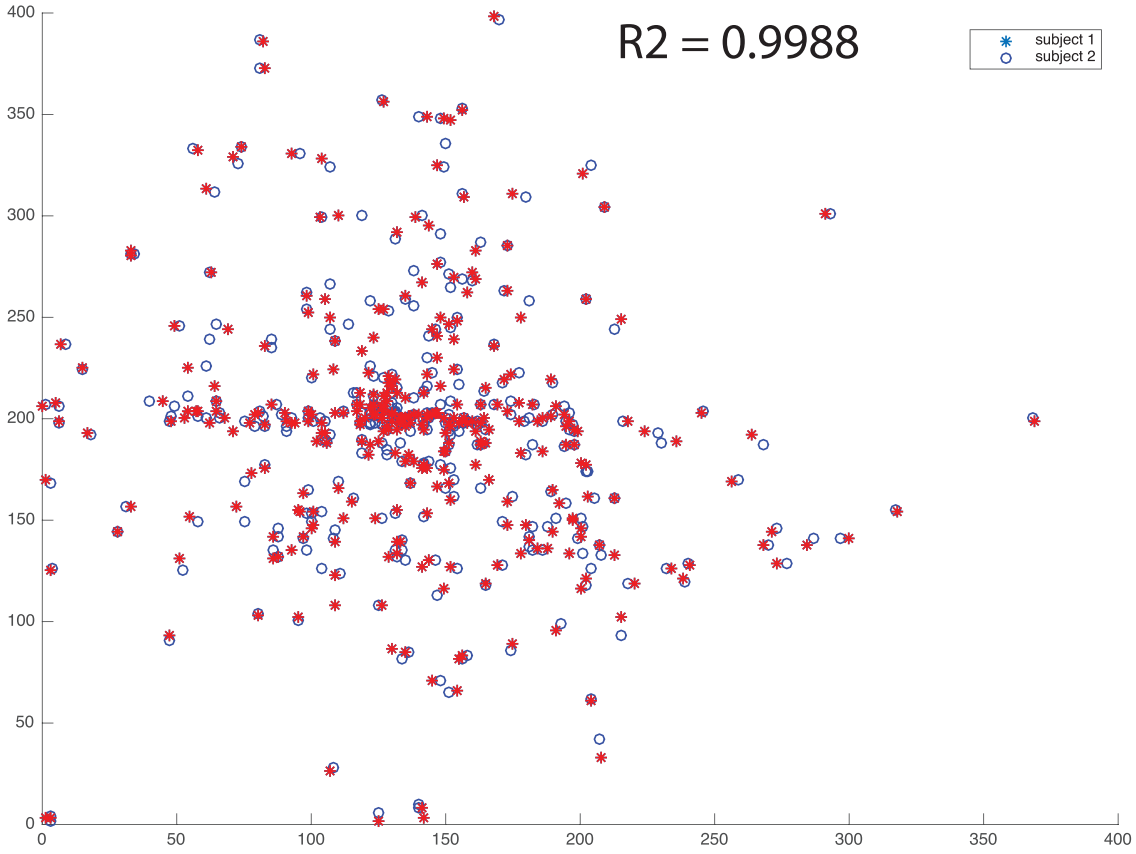
Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*, 787.

Zhang, J., & Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 153–160). IEEE.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7):32, 1–20, doi:10.1167/8.7.32. [PubMed] [Article]

Appendix A

Figure 12 shows the consistency results of our two annotators in marking the vanishing point location. Each point is the location of an annotated VP in the image. R^2 is the correlation coefficient of the 2D vectors. The bar chart shows the distribution of annotation differences between the two subjects. For about 280 of the images (out of 319), the difference is smaller than or equal to 5 pixels on a 640 × 480 pixel image. Since annotators were very consistent, we used annotations of only one of them in our analyses.



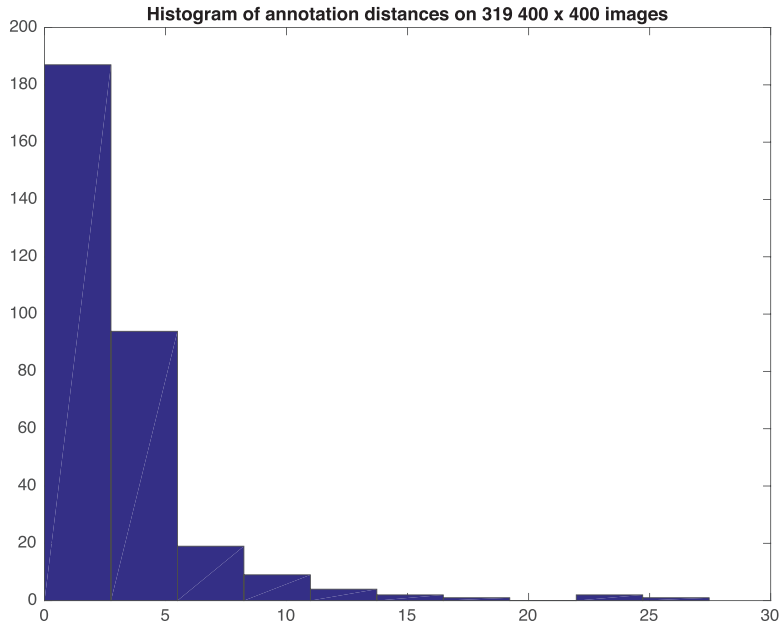


Figure 12. (A) Annotated VP locations by two annotators. (B) Histogram of vanishing point differences.

Appendix B

Figure 13 shows the optimization of the VP sigma parameter (left) and sAUC scores of Model + VP versus VP for three saliency models (Right). See also the section entitled "Addressing center-bias."

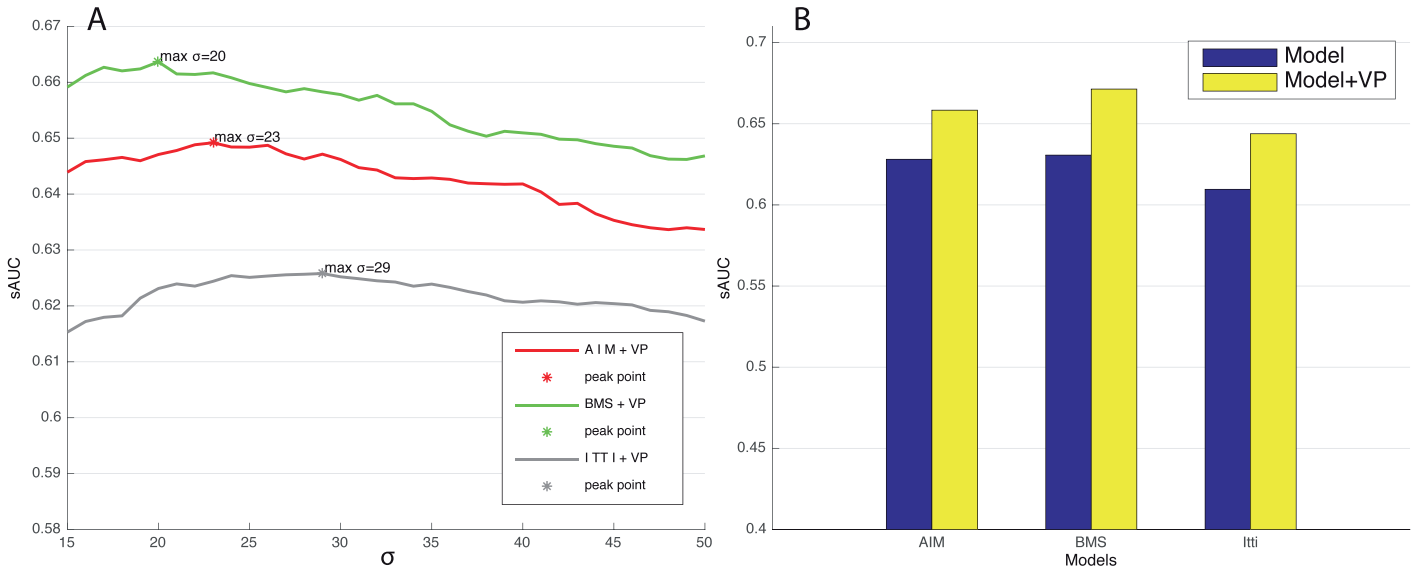


Figure 13. (A) Optimizing the M+VP model for VP sigma over the training set containing 50 images. (B) sAUC score of the learned model with the best VP sigma over 269 test images. Note that the M + VP model significantly outperforms the M model using three base saliency models.

Appendix C

Figure 14 shows the response time of individual subjects in the visual search experiment.

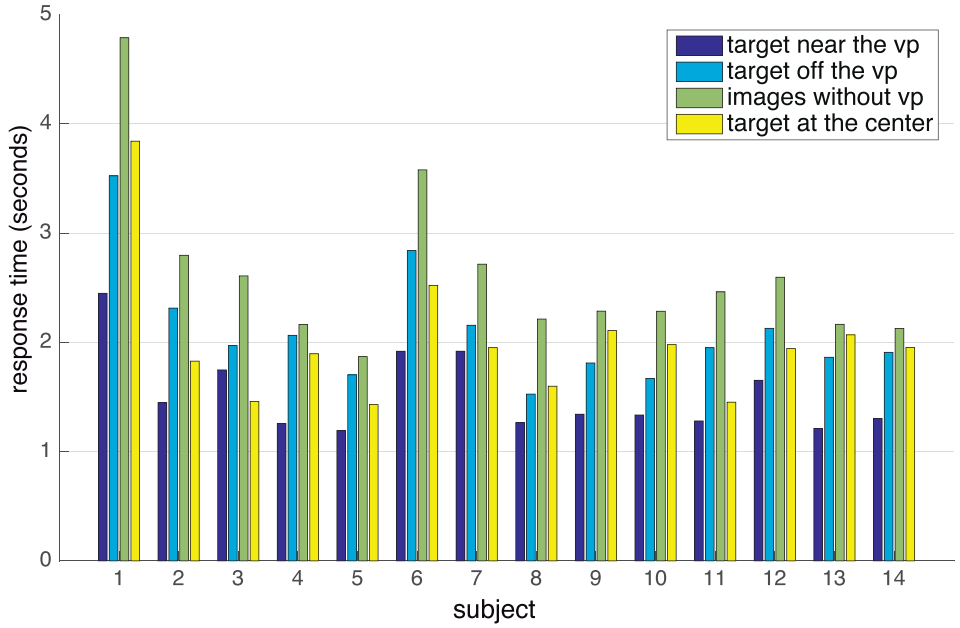


Figure 14. Response times of subjects in the visual search experiment.

Appendix D

A convolutional neural network (Krizhevsky et al., 2012) was trained to predict whether a vanishing point exists on a scene or not. Positive images were road scenes downloaded from YouTube® (37,497 images), and negative images were randomly selected scenes from the Web (32,419 images). The network's accuracy was 98.9% (chance = 50%). Interestingly, we observed that some vanishing-point-selective neurons emerged in higher convolutional layers (Conv5 layer) of the network. Figure 15 shows some of those neurons along with their most favorite stimuli. Please see Borji (2016) for more details.

Figure 15. Vanishing-point-selective neurons emerged in Conv5 layer of the Alexnet after training on the vanishing point existence prediction task.