

Fixation prediction with a combined model of bottom-up saliency and vanishing point

Mengyang Feng⁺ Ali Borji[†] Huchuan Lu⁺

[†]Computer Science Department, University of Wisconsin - Milwaukee, USA

⁺Department of Electrical Engineering, Dalian University of Technology, China
 archerfmy@163.com, borji@uwm.edu, luhuchuan@gmail.com

Abstract

By predicting where humans look in natural scenes, we can understand how they perceive complex natural scenes and prioritize information for further high-level visual processing. Several models have been proposed for this purpose, yet there is a gap between best existing saliency models and human performance. While many researchers have developed purely computational models for fixation prediction, less attempts have been made to discover cognitive factors that guide gaze. Here, we study the effect of a particular type of scene structural information, known as the vanishing point, and show that human gaze is attracted to the vanishing point regions. We record eye movements of 10 observers over 532 images, out of which 319 have vanishing points. We then construct a combined model of traditional saliency and a vanishing point channel and show that our model outperforms state of the art saliency models using three scores on our dataset.

1. Introduction

Visual attention and eye movements are crucial in understanding complex scenes. Primates use focal visual attention and rapid eye movements to analyze complex visual inputs in real-time, in a manner that highly depends on current behavioral priorities and goals. Yarbus (1967) [31] demonstrated a striking example of how a verbally-communicated task specification may dramatically affect attentional deployment and eye movements. He argued that variable spatiotemporal characteristics of scanpath for different task specifications exemplify the extent to which behavioral goals may affect eye movements and scene analysis. Some eye tracking experiments in the context of spoken sentence comprehension have shown that the interplay between task demands and active vision is reciprocal. For example, Tanenhaus et al. (1995) [26] tracked fixations of subjects when they received ambiguous verbal instructions regarding manipulating objects on a table. Tanenhaus et al. demonstrated that visual context influenced spoken word recogni-

tion 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.

Following two seminal works, Feature Integration Theory by Triesman and Gelade [29] and the computational attention architecture by Koch and Ullman [15], several attention models have been proposed to detect bottom-up salient regions that stand out from their surroundings in an image [1, 3, 13]. These models can be classified under three categories: 1) purely computational, 2) purely cognitive, and 3) a hybrid of both computational and cognitive. Models in the first category intend to detect salient regions often by using machine learning or statistical tools. For example, some researchers have formulated the problem as a classification problem by trying to estimate which points (and to what degree) will be looked at by humans (e.g., [9, 10, 32, 14, 4]). Some studies in the second category have been investigating cognitive factors that influence eye movements in free viewing of natural scenes. These are often behavioral studies that accurately formulate/analyse hypotheses and rule out confounding factors. For example, it has been shown that eye movements are driven to the center of objects [21] and scenes [27] or gaze direction of actors in scenes direct viewers' gaze [2, 23]. Models in the third category are either inspired by the mechanisms of human attention and mimic it (e.g., [12, 8]) and/or have used a set of cognitive factors to build a model to predict fixations (e.g., [6, 22]). Note that some studies and models fall under more than one category and the categories are not exclusive.

Several cognitive cues that attract attention and guide eye movements have already been discovered (e.g., color, texture, motion, text [6], face [6], object center-bias [21], scene center-bias [27], cultural cues [7, 25], and gaze direction [5, 2]). Scene structural information such as scene gist (global context), scene layout, horizontal line, depth, and openness influence eye movements as well as human scene categorization [28]. Here, we systematically investigate the role of a particular type of scene structural informa-



Figure 1. Example stimuli with vanishing points (yellow boxes) and fixations (dots). For some images with highly salient items, the vanishing point attracts less attention (e.g., bottom-right image). For some images, salient content happens at the vanishing point while for some others it does not.

tion, known as the vanishing point (VP) and perspective, on eye movements in free-viewing of pictures of natural scenes and propose a combined model of bottom-up saliency and VP. Therefore, our model is classified under the third category mentioned above.

In graphical perspective, a vanishing point is a 2D point (in image plane) which is the intersection of parallel lines in the 3D world (but not parallel to the image plane). VP can be seen in fields, rail roads, streets, tunnels, forest, buildings, objects such as ladder, etc. It has been used in camera calibration, 3D reconstruction as well as in painting.

2. Data Collection

2.1. Stimuli

Our stimuli contains 319 color images with vanishing points with resolution of 1920×1080 pixels (with added gray margins while preserving the aspect ratio) from different categories. Firstly, we collected 700 images from Google search, MIT300 [13] and DUT-OMRON [30] datasets. We ruled out images with more than one vanishing point and images with complex texture informations which may cost the disadvantage of (automatically) detecting the vanishing point. Eventually, we were left with 319 images and manually annotated vanishing points by drawing rectangles around them. Two members of our laboratory completed the annotation task together. Average height of VP rectangles is 10px and average width is 14px (only center of VP is used here). Since showing only images with a vanishing point may generate a viewing bias in observers and draw them automatically to vanishing points, we then gathered additional 213 images without vanishing points, and shuffled them among images with VPs. Therefore, viewers would not know in advance whether a presented image will have a VP. We had 532 images in total to record human fixations, out of which 319 had VP and 213 did not. In our modeling and experiments here, we only analyze 319 images with VPs. Figure 1 shows examples of our stimuli, labeled vanishing points, as well as fixation locations.

Figure 2.A shows average VP annotation map as well as average fixation map over 319 images with VPs. Both of these maps have maxima at the image center making center-bias a potential confounding factor which we will address extensively in our analyses. This figure also shows the histogram of VP window size. 82.25% of VP rectangles have size smaller than 0.2% the image area.

2.2. Eye tracking

Observers: We had 10 observers (6 male, 4 female) in total. Mean observer age was 22 (min=21, max=24, median 22, std 0.84). Observers were undergraduates at our university from different majors and cities. Observers had normal or corrected-to-normal vision and received course credit for participation. They were naive to the purpose of the experiment and had not previously seen the stimuli.

Procedure: Following the fixation cross, a target image was shown for 4 seconds followed by 3 seconds gray screen. Observers sat 60 cm away from a 19 inch LCD monitor screen such that scenes subtended approximately $37.6 \text{ degree} \times 24 \text{ degrees}$ of visual angle. A chin rest was used to stabilize head movements. Stimuli were presented at 60Hz 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 1000Hz. The eye tracker was calibrated using 5 points calibration at the beginning of each recording session. Observers viewed images in random order.

3. Our Model

In this section, we present details of our learning model with human annotations first. We then mention how we will automatically detect VPs to replace human annotations. We also compare performance of our model with human annotations and with automatic detections.

3.1. Learning a combined saliency map

Each image pixel is represented by $X = [s \ v]$ where s is the output of a bottom-up saliency model (e.g., AIM [4], BMS [32], and Itti [12]). v is the value from the vanishing point map (VP) modeled as a variable size Gaussian placed at the vanishing point as shown in Figures 2.B & 2.C¹:

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

where (i, j) is the coordinate of the annotated vanishing point and σ_{vp} is the (variable) standard deviation of the Gaussian blob. In section 3.2, the coordinate of the vanishing point will be replaced by our automatic VP detector.

¹We experimentally found that the Gaussian form of VP works better than a rectangle or a circle.

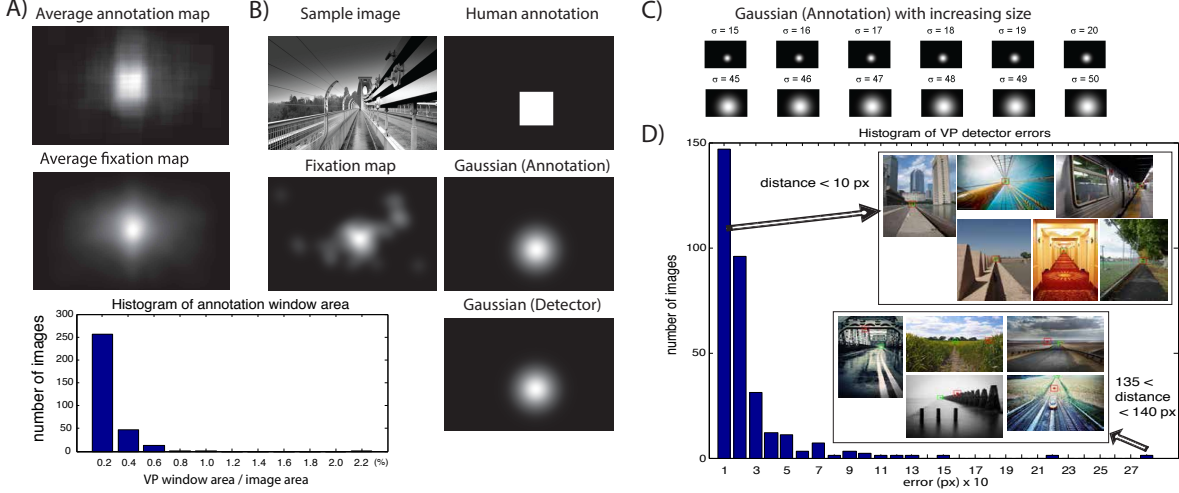


Figure 2. A) Average VP map and average fixation map over images with vanishing points. B) A sample image, its vanishing point map, eye movements, Gaussian blob placed at the vanishing point, and Gaussian placed at the vanishing point using our automatic VP detector. C) Illustration of the Gaussian at the vanishing point with increasing σ_{vp} . D) Histogram of vanishing point detector errors. Bins here are $[0,5)$, $[5, 10)$, \dots , and $[135,140)$. Some examples when detector works well or fails are also shown (Green= Ground truth, Red= Detection).

We aim to learn $f(X) = W^T X + b$ which is a binary function determining whether location with feature vector X should be attended or not. To do so, we use a SVM with a linear kernel. For a test pixel, we assign the $m = W^T X + b$ as the label of it. Final saliency values are then normalized for each map (i.e., $(m - \min) / (\max - \min)$). We avoid using complicated non-linear learning functions (e.g., boosting) here deliberately, since we are interested to find out the real added value of the vanishing point.

We choose 50 random images for training the SVM and the rest 269 images for testing. 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., 5000 samples in total. Note that we cut off the added gray margins and resized the maximum length of the image side to 400 pixels while preserving the aspect ratio (to reduce the calculation).

We learn the combined models (e.g., AIM + VP, BMS + VP, and Itti + VP) and compare them with the original bottom-up saliency models, respectively.

3.2. Automatic detection of vanishing points

Several methods for detecting vanishing points in an image exist (See [16]). Some methods utilize line segments detected in an image. Some other approaches consider intensity gradients of the image pixel. There can be several vanishing points present in an image. Here, our aim is to detect the vanishing point that corresponds to the principal directions (lines) in a scene.

Our method also utilizes line segments to get the vanishing points. For an input image, we use the PB boundary detection algorithm [20] to obtain the boundary map B .

$B(i, j)$ gives the probability of a boundary at each pixel (i, j) . We then applied Hough Transform [11] to detect line segments. Since the input of the Hough Transform should be a binary map, we turn B into a binary map using an adaptive threshold,

$$B_2(i, j) = \begin{cases} 1 & B(i, j) \geq t \\ 0 & B(i, j) < t \end{cases} \quad (2)$$

where $t = 10 \times \frac{\sum B(i, j)}{\text{height} \times \text{width}}$. In this work, height and width are the size of the B map. t is chosen by experience. Then the line segments map L is computed as $L = \text{Hough}(B_2, \theta, lt)$ where θ is the angle of lines which could be detected, and lt represents the threshold when choosing a line. In this work, we set $\theta = 180^\circ$ that means lines from every direction can be detected. And $lt = 60$, which means that lines which have more than 60 pixels on them can be detected. Note that, the parameters t and lt ensure that only the large line segments could be detected.

$$L(i, j) = \begin{cases} 1, & (i, j) \in l_d \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where l_d presents the detected lines. From the line segments map $L(i, j)$, we can get the intersections of those lines. Our aim is to detect the vanishing point that corresponds to the principal directions in a scene. So, we can assume that the location (X, Y) where most intersections happen around it could be the vanishing point. More specifically, if two intersections' Euclidean distance is smaller than 10 pixels, we consider that they are the neighbor points, and calculate the number of neighboring points around each intersection. Then, we can find the intersection (i_v, j_v) which has most neighboring points around and calculate the VP

location (X, Y) using this formula,

$$X = \frac{\sum_1^M x_i}{M}; Y = \frac{\sum_1^M y_i}{M}; i = 1, 2, \dots, M \quad (4)$$

where M is the number of the neighbor points around (i_v, j_v) , and (x_i, y_i) presents the coordinate of the i th neighbor. The method performed well over our dataset.

4. Experiments and Results

Firstly, we aim to optimize our combined model by finding the best σ_{vp} . Table 1 summarizes the results by reporting the point where performance is maximum. σ_{vp} is changing from $\sigma_{vp} = 15$ pixel to $\sigma_{vp} = 50$ pixel. We observe more than 10% improvement of Model + VP versus Model using the AUC scores with any of the three models. Improvement using NSS is more than 50% while improvement using CC is more than 60%. Considering both NSS and CC scores, we determined the σ_{vp}^{best} by adding the normalized NSS and CC scores and selecting the σ_{vp} value corresponding to the peak. Then Model + VP_b represents the optimized combined model.

Since center-bias [27] is an important confounding factor, here we compare the Model + VP_b + CG(Central Gaussian map) to Model + CG to see whether the VP is the main cause or not. Figure 3 shows scores of models as a function of σ_{cg} (i.e., the σ of the Central Gaussian). As this figure shows increasing σ_{cg} increases the AUC score until it saturates. Performance peaks using NSS [24] and CC and then declines. Model + CG works better than CG and VP only maps but performs below Model + VP_b + CG. This trend happens using all three scores but is more prevalent using CC and NSS scores. Interestingly, performance using our VP detector is very close to the performance using human annotations (although slightly lower). The automatic VP detector has error in locating the correct location of the VP for some images, but this error is negligible because it does not affect the placement of the Gaussian blob (i.e., smoothing). In both rows, baseline models score below all shown models including VP only and CG only models.

To investigate the effect of center-bias on our results we have conducted two analyses shown in Figure 4. Figure 4.A shows performance of the center-bias modulated VP-added model versus center-bias modulated model for each image. For 221 of the images, we observe an improvement of the former over the latter. In fact, when observing these 221 images, we found that for the majority of the images, VP happened off the center. This means that adding VP increases the results additional to what adding center-bias offers. In Figure 4.B, we plot the AUC score of the vanishing point (VP) map versus the Central Gaussian map (CG). As this figure shows, VP map wins over the Central Gaussian map for some images (36.1% of test images). For many other

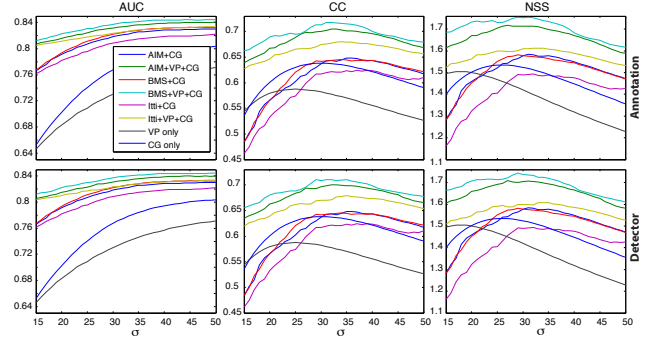


Figure 3. Performance of models using AUC, CC, and NSS scores as a function of σ_{cg} in Model + VP_b + CG. The top row shows model performance using the human annotation and the bottom row shows the performance using the automatic VP detector. As you can see performance using detector and human annotations are very close to each other. Since the detector error is small relative to Gaussian size, it gets canceled.

images, VP happens at the image center. Thus, we conclude that vanishing point and central bias are two different phenomena with distinct effects, although over our stimuli they coincide in many images by construction.

Tables 2 compares Model + CG versus Model + CG + VP_b which addresses center-bias. Improvement of Model + VP_b + CG over Model + CG is smaller using 3 scores (about 2.5% using AUC average over 3 models, about 9.1% using NSS, and about 9.7% using CC). Investigating the parameters of the discriminant line learned by SVM (i.e., weight of the BU and VP) shows that both baseline model and VP map are involved in the final combination.

To check the statistical significance of our results, we perform cross validation by randomly splitting data into two parts (50 train and 269 test). We train our SVM model on the train set and apply it to the test set. We repeat this proce-

Score		AIM	AIM + VP	BMS	BMS + VP	Itti	Itti + VP	VP only
AUC	A	0.719	0.807 (50)	0.708	0.812 (50)	0.727	0.801 (50)	0.771 (50)
	D	0.719	0.804 (50)	0.708	0.809 (50)	0.727	0.798 (50)	0.771 (50)
	I	-	12.2%	-	14.7%	-	10.2%	-
	W	-	[5.8, 6.5]	-	[7.1, 6.1]	-	[5.1, 6.0]	-
NSS	A	0.957	1.495 (31)	0.916	1.544 (26)	0.903	1.412 (37)	1.505 (18)
	D	0.957	1.483 (27)	0.916	1.512 (26)	0.903	1.397 (37)	1.505 (18)
	I	-	56.2%	-	68.6%	-	56.4%	-
	W	-	[5.8, 6.5]	-	[7.1, 6.1]	-	[5.1, 6.0]	-
CC	A	0.356	0.601 (38)	0.340	0.612 (33)	0.362	0.585 (37)	0.587 (25)
	D	0.356	0.591 (36)	0.340	0.600 (33)	0.362	0.578 (37)	0.587 (25)
	I	-	68.6%	-	80%	-	61.6%	-
	W	-	[8.0, 6.9]	-	[9.5, 6.7]	-	[6.9, 6.9]	-

Table 1. Scores of our combined model (Model + VP) vs. the original model and VP only channel. Numbers in parentheses are the best σ_{vp} where performance peaks using different scores. Improvement is measured with respect to the original model. W is the parameters of the line learned by the SVM classifier. σ_{vp}^{best} using human annotations and auto detector for AIM, BMS, and Itti models in order are 31, 27, and 37 pixels. (A= Annotation, D= Detection, I= Improvement, W= Weights).

Score		AIM+C	AIM+C+V _b	BMS+C	BMS+C+V _b	Itti+C	Itti+C+V _b	C only
AUC	Annotation (A)	0.803 (48)	0.841 (50)	0.833 (45)	0.845 (45)	0.822 (50)	0.834 (50)	0.804 (50)
	Detection (D)	0.803 (48)	0.840 (45)	0.833 (45)	0.845 (49)	0.822 (50)	0.834 (49)	0.804 (50)
	Improvement (I)	-	4.7%	-	1.4%	-	1.5%	-
	Weights (W)	[4.5,4.9]	[4.1,2.0,4.6]	[5.9,5.4]	[5.5,2.3,4.8]	[4.1, 5.4]	[3.7,2.5,4.9]	-
NSS	A	1.584(31)	1.719(27)	1.582(29)	1.755(29)	1.493(32)	1.613(34)	1.535(28)
	D	1.584(31)	1.709(27)	1.582(29)	1.745(29)	1.493(32)	1.608(35)	1.535(25)
	I	-	8.5%	-	10.9%	-	8.0%	-
	W	[7.7, 6.9]	[7.2, 4.3,5.3]	[9.2,7.2]	[8.2,4.3,5.2]	[6.2,6]	[4.9,4.2,4.8]	-
CC	A	0.648(35)	0.705(33)	0.645(34)	0.718(32)	0.624(37)	0.680(34)	0.638(30)
	D	0.648(35)	0.700(33)	0.645(34)	0.711(32)	0.624(37)	0.679(35)	0.638(30)
	I	-	8.8%	-	11.3%	-	9.0%	-
	W	[6.6,6.5]	[6.1,3.6,5.4]	[8.2,6.3]	[7.3,3.9,5.5]	[5.6,5.7]	[4.9,4.2,4.8]	-

Table 2. Performance of models using AUC, CC, and NSS scores as a function of σ_{cg} in Model + VP_b + CG.

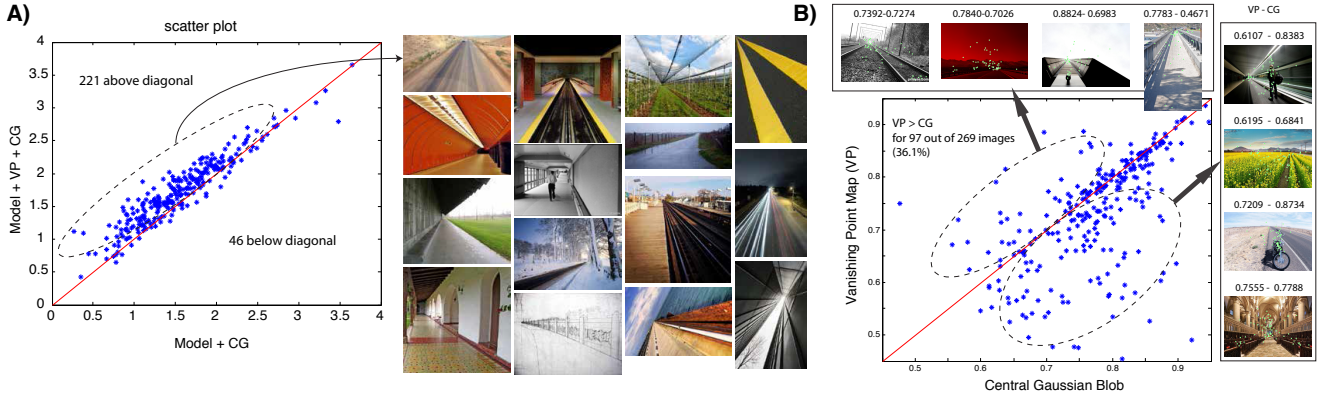


Figure 4. A) Scatter plot of Model + CG versus Model + VP + CG using NSS score. Each dot represents one image. CG stands for Central Gaussian. This plot shows the added value of VP over the original model taking into account the center bias confound. So we can make sure that added value is not because of the center-bias. For 221 of images, we observe an improvement. Vanishing points usually happens off center on these images. We did the same analysis by plotting the Model + VP versus Model and observed that for 243 of images performance is improved. B) AUC score of the VP map versus Central Gaussian map for $\sigma_{vp} = \sigma_{cg} = 31$ pixels. Some examples above/below diagonal are shown.

Score	Model	M + VP _b + CG _b vs. M + CG _b	M + VP _b vs. VP _b	M + VP _b vs. M	VP _b vs. Chance
AUC	AIM	0.833 vs. 0.819 p=1.574e-24	0.793 vs. 0.739 p=6.866e-30	0.793 vs. 0.720 p=4.193e-35	0.739 vs. 0.5 p=4.905e-40
	BMS	0.837 vs. 0.823 p=3.993e-21	0.798 vs. 0.719 p=4.466e-32	0.798 vs. 0.711 p=6.759e-34	0.719 vs. 0.5 p=1.161e-39
	Itti	0.826 vs. 0.811 p=5.332e-21	0.792 vs. 0.756 p=2.474e-22	0.792 vs. 0.730 p=1.075e-30	0.756 vs. 0.5 p=5.840e-40
NSS	AIM	1.695 vs. 1.545 p=4.694e-21	1.467 vs. 1.450 p=1.932e-03	1.467 vs. 0.953 p=1.704e-28	1.450 vs. 0 p=6.875e-39
	BMS	1.751 vs. 1.587 p=7.277e-22	1.543 vs. 1.508 p=5.273e-06	1.543 vs. 0.916 p=6.738e-31	1.508 vs. 0 p=4.217e-38
	Itti	1.592 vs. 1.445 p=1.384e-17	1.361 vs. 1.381 p=7.499e-03	1.361 vs. 0.918 p=2.091e-25	1.381 vs. 0 p=1.370e-39
CC	AIM	0.697 vs. 0.628 p=6.328e-20	0.584 vs. 0.598 p=4.755e-06	0.584 vs. 0.358 p=5.343e-28	0.598 -
	BMS	0.720 vs. 0.652 p=4.673e-21	0.609 vs. 0.608 p=8.347e-01	0.609 vs. 0.341 p=2.709e-30	0.608 -
	Itti	0.672 vs. 0.603 p=2.279e-17	0.563 vs. 0.580 p=3.256e-05	0.563 vs. 0.369 p=3.823e-25	0.580 -

Table 3. Statistical analysis of results and model comparison.

procedure 20 times and compare the means. Results of statistical tests using t-test are shown in Table 3. We perform four comparisons: 1) M + VP_b + CG_b(Central Gaussian using optimized σ_{cg} , similar to the σ_{vp}^{best}) vs. M + CG_b, 2) M +

VP_b vs. VP_b, 3) M + VP_b vs. M, and 4) VP_b vs. Chance. We find that VP model performs significantly above the chance level in predicting fixations. Adding VP to models significantly outperforms both VP and baseline models (taken individually). Tackling center-bias, by adding CG_b to Model and to Model + VP_b, shows that VP is a significant cue in guiding gaze, independent of center-bias. We obtain the same pattern of results using NSS and CC scores.

Figure 5 shows examples where our combined model fails (i.e., M + VP_b scores lower than M). In almost all these cases, an object off the vanishing point overrides the VP effect. Figure 6 shows examples where our model performs well along with scores of models.

5. Discussion and Conclusion

We showed that vanishing point is a strong predictor of fixations in free viewing task and proposed a combined model of bottom-up saliency (using three state of the art models) and VP. Our model outperforms baseline models significantly with and without center-bias using three scores. We also showed that VP map performs significantly above chance. Since VP happens commonly in real life

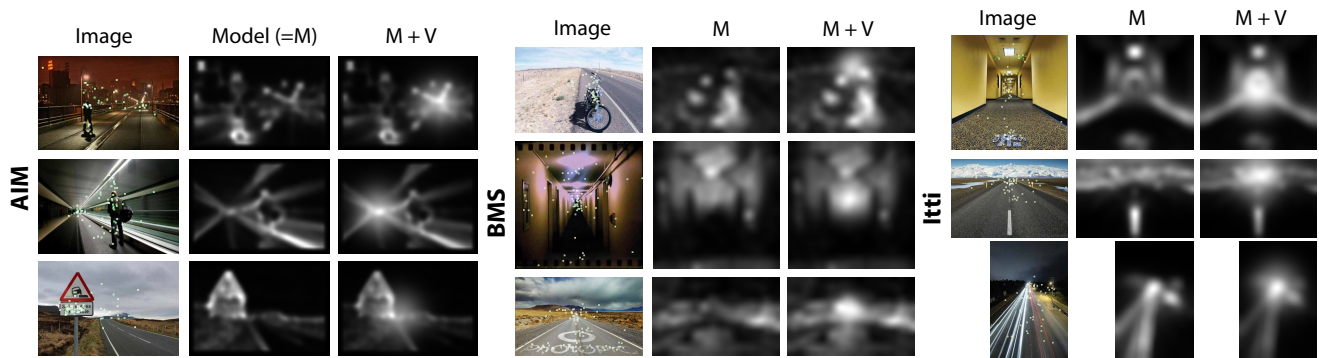


Figure 5. Cases where our combined model performed poorly. In almost all of these cases, an object off the vanishing point overrides the VP and attract fixations.

when taking pictures, we believe that adding it to models can in general enhance fixation prediction power.

We intend to study the followings in future: 1) Whether (and to what extent) people prioritize vanishing points in presence of other salient cues in a scene? 2) Here, we added VP channel to images with a vanishing point. While this was not a problem with annotations, ultimately, we would like to add VP to only those images which have VP. For this we should automatically decide whether an image has VP or not (i.e., How much false positives of our detector will hurt?). In this regard, we will also consider images with multiple VPs, 3) We aim to relate our findings to other cues that might influence fixations in a similar fashion (but independently), for cues such as "focus of expansion" or "tangent line" [17, 18] and 4) we will consider other methods for detecting vanishing points in images (e.g., using convolutional neural networks [19]).

We will share our dataset for further investigation of the role of the vanishing point cue in guiding gaze in free viewing. Hopefully our work will encourage more research toward discovering behavioral cues that guide attention and gaze in spatial and spatio-temporal domains.

References

- [1] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *35(1)*:185–207, 2013.
- [2] A. Borji, D. Parks, and L. Itti. Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of vision*, 14(13):3, 2014.
- [3] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.
- [4] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2005.
- [5] M. S. Castelhan, M. Wieth, and J. M. Henderson. 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*, pages 251–262. Springer, 2007.
- [6] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *NIPS*, 2007.
- [7] H. F. Chua, J. E. Boland, and R. E. Nisbett. Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12629–12633, 2005.
- [8] A. Garcia-Diaz, V. Leboran, X. R. Fdez-Vidal, and X. M. Pardo. On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision.*, 12(6), 2012.
- [9] J. Harel, C. Koch, P. Perona, et al. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545, 2007.
- [10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8. IEEE, 2007.
- [11] P. V. Hough. Machine analysis of bubble chamber pictures. In *International Conference on High Energy Accelerators and Instrumentation*, volume 73, 1959.
- [12] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE. PAMI*, 1998.
- [13] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.
- [14] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *CVPR, 2009*, pages 2106–2113, 2009.
- [15] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.
- [16] H. Kong, J.-Y. Audibert, and J. Ponce. Vanishing point detection for road detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 96–103. IEEE, 2009.
- [17] M. F. Land and D. N. Lee. Where do we look when we steer. *Nature*, 1994.
- [18] M. F. Land and B. W. Tatler. Steering with the head: The visual strategy of a racing driver. *Current Biology*, 11(15):1215–1220, 2001.
- [19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [20] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color,

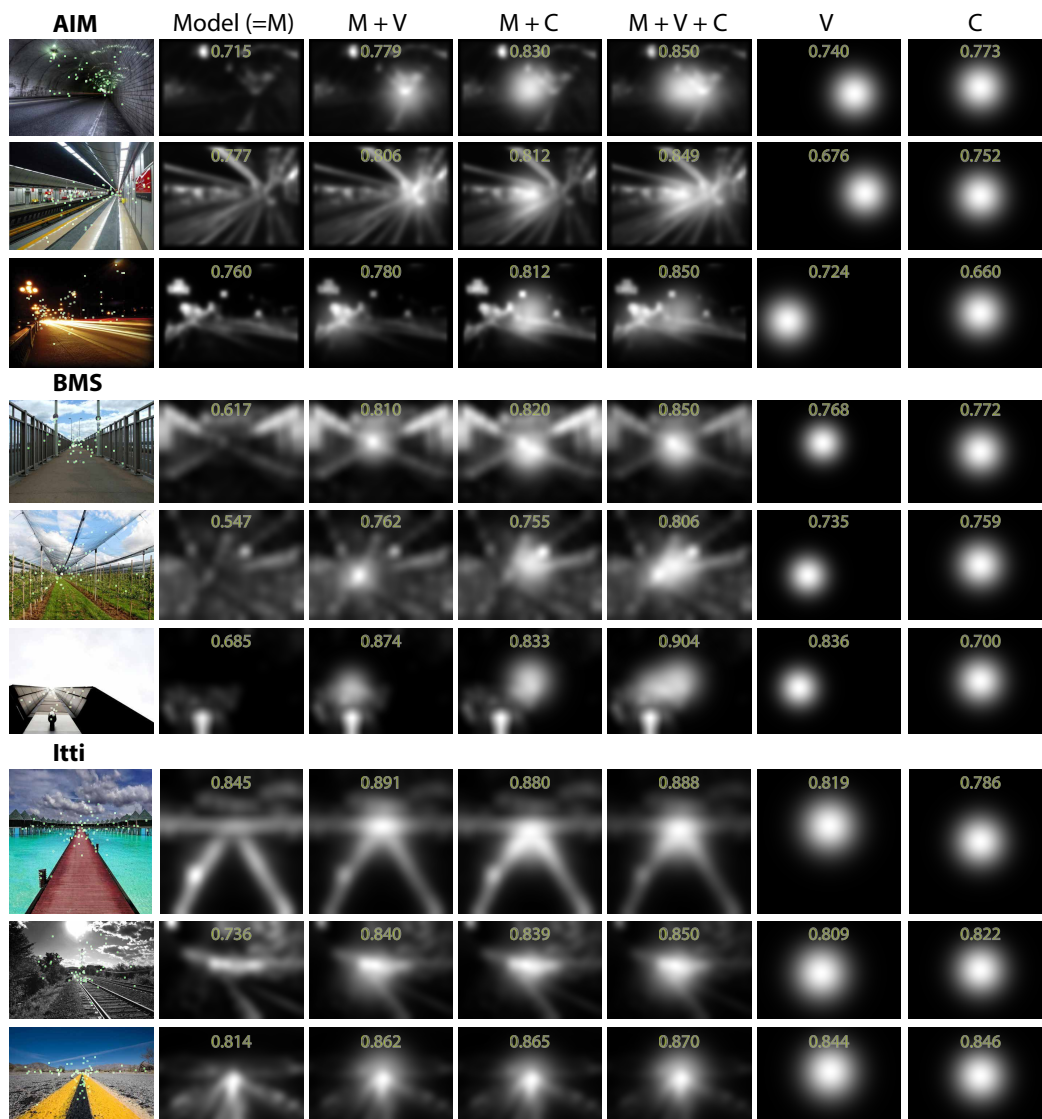


Figure 6. Cases where our combined model performed well (using AUC score).

- and texture cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(5):530–549, 2004.
- [21] A. Nuthmann and J. M. Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8):20, 2010.
- [22] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.
- [23] D. Parks, A. Borji, and L. Itti. Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes. *Vision research*, 2014.
- [24] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*.
- [25] K. Rayner, M. S. Castelhana, and J. Yang. Eye movements when looking at unusual/weird scenes: Are there cultural differences? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1):254, 2009.
- [26] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995.
- [27] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision research*, 45(5):643–659, 2005.
- [28] A. Torralba, A. Oliva, M. S. Castelhana, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.
- [29] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.
- [30] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*. IEEE, 2013.
- [31] A. L. Yarbus, B. Haigh, and L. A. Riggs. *Eye movements and vision*, volume 2. Plenum press New York, 1967.
- [32] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *CVPR*, pages 153–160. IEEE, 2013.