Visual saliency predicts gaze during real-world driving task

Richard Veale (veale.richard.7c@kyoto-u.ac.jp)
Graduate School of Medicine, Kyoto University, Kyoto, Japan

Kenji Murase (murase.ke@mazda.co.jp)
Technical Research Center, Mazda Motor Co., Hiroshima, Japan

Masayuki Watanabe (watanabe.masay@mazda.co.jp)
Technical Research Center, Mazda Motor Co., Hiroshima, Japan

Tadashi Isa (isa.tadashi.7u@kyoto-u.ac.jp)
Graduate School of Medicine, Kyoto University, Kyoto, Japan

Abstract

Models of bottom-up visual attention such as the "saliency map" predict overt gaze under laboratory conditions while subjects view static images or videos while seated. Here, we show that the saliency map model predicts gaze at similar rates even when applied to video from a head-camera as part of a wearable eye-tracking system (Tobii Pro Glasses 2) while subjects drive an automobile or are passively driven while sitting in the front passenger-side seat. The ability of saliency to predict gaze varies depending on the driving task (saliency better predicts passenger gaze) and external conditions (saliency better predicts gaze at night). We further demonstrate that predictive performance is improved when the head-camera video is transformed to retinal coordinates before feeding it to the saliency model.

Keywords: Visual Saliency; Wearable Eyetracker; Saliency Map; Bottom-up Visual Attention

Introduction

Models of bottom-up visual attention such as the "saliency map" (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000) predict overt gaze in humans freely viewing pictures (Bylinskii, Judd, Oliva, Torralba, & Durand, 2018). While saliency predicts looking even to dynamic video stimuli in humans and other primates (Chen et al., 2021), previous research has generally assumed that saliency is applied to a "visual stimulus" (an image or video) which the subject is viewing on a fixed screen. Under such conditions, different parts of the visual stimulus compete to pull attention, and this is what visual attention models predict. In contrast, under natural conditions, the visual stimulus is the subject's entire visual field (and possibly things outside the visual field, e.g. visual memories). Under unconstrained natural conditions, subjects execute complex gaze movements involving head, eye, and body movements, to accomplish arbitrary and often implicit ecologically relevant tasks, such as satisfying curiosity, searching for food, appreciating beautiful vistas, navigating in the environment, or hitting a baseball (Land, 2015, 2009). Even under free-viewing conditions with simple (underconstrained) instructions such as "watch the video", subject behavior reflects unknown and idiosyncratic self-generated tasks depending on the mood of the subject, his recent memories, and his personality and preferences (Borji & Itti, 2014; Koehler, Guo, Zhang, & Eckstein, 2014).

We address the question of whether saliency map models can successfully predict gaze using input from a head-fixed camera (also known as "egocentric" view) even under real-world and freely-moving conditions. Furthermore, we address what kinds of preprocessing steps or modifications to the model are necessary to ensure proper function. These preprocessing steps address our uncertainty regarding the proper coordinate system of bottom-up visual attention. Is the visual input to attention best presented in retinotopic (eye-centered), egocentric (head-centered or torso-centered), or allocentric (world-centered) coordinates? Furthermore, should visual input be degraded to match the reduction in visual acuity that occurs at higher retinotopic eccentricities ("foveation")? It is known that the instantaneous drive for (voluntary) gaze shifts is driven by cells in the deeper layers of midbrain superior colliculus (SC), which represents both visual targets as well as gaze targets in a retinotopic coordinate system (in other words, centered on the retina) (Takahashi & Veale, 2023).

In this paper, we use a sedentary task involving little body movement (driving a car, or riding in the passenger seat of a car). Driving requires heavy concentration (looking where the car is going, scanning for dangers, checking the rearview mirrors and blind spots (Lappi, 2022)). Estimating driver attention is not a new field. A plethora of attempted models including saliency have been applied to datasets (usually collected in simulators) (Kotseruba & Tsotsos, 2022). While driving, subject gaze behavior likely differs significantly from "free viewing" conditions in the laboratory (Kübler et al., 2015; Lappi, 2016). The nature of driving requires subjects to look "straight ahead" often (towards the vanishing point) (Palazzi, Abati, Solera, Cucchiara, et al., 2018).

We compare the performance of a model of bottom-up visual attention (the saliency map model) under various experimental conditions (driver versus passenger, night versus day), as well as various stimulus preprocessing conditions (head-centered versus eye-centered, foveated versus unfoveated). Previous models of visual attention while driving (e.g. (Palazzi et al., 2018)) have implicitly assumed that attention is described in "car-coordinates" (i.e. from the point of view of a fixed forward-facing camera attached to the top of the car), without describing how an attention model might handle other coordinate systems or whether performance is improved or reduced by transformations of the visual input. (Adeli, Vitu, & Zelinsky, 2017) showed that transformation

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

of the visual stimuli into SC-coordinates (Ottes, Van Gisbergen, & Eggermont, 1986) improved the performance of a saliency map model in predicting the endpoints of saccades during free viewing or search tasks, although they included several additional steps in the model. Other research into egocentric visual attention (Schumann et al., 2008) shows that while human head and eye positions are aligned, salient (high-contrast, high-entropy) regions tend to be centered in retinal coordinates but not necessarily in head coordinates. However, acute neural recordings from monkey hippocampus and entorhinal cortex, related to e.g. place cells described in rodent, have confirmed that mental-map orientation (for e.g. navigation) is primarily couched in terms of head-direction in contrast to torso-direction or eye-direction (Mao, 2023; Mao et al., 2021), implying that spatial decisions such as overt attention may occur in a head-centered coordinate system, even if the final output for gaze is achieved retinotopically (Takahashi & Veale, 2023). However, the visual information available to the brain arrives with self-motion and blur removed by the counter-rotation of the eyes caused by the vestibulo-ocular reflex (VOR), suggesting that gaze targets are not influenced by perceived motion caused by self-movement.

Previous approaches to gaze prediction of egocentric video have all use use head-centered video (albeit with deep neural networks which apply unknown transformations of the input). Several studies have investigated the application of saliency to predicting gaze of a wearable eyetracker, beginning with (Yamada et al., 2012). The field of egocentric gaze prediction has advanced significantly, albeit with a focus on integrating additional cues such as hand location (detected in the subject's own egocentric view) to predict gaze better, or to identify tasks, which are highly predictive of gaze. The additional cues are combined with visual information to predict gaze (Huang, Cai, Li, Lu, & Sato, 2020). (Tavakoli, Rahtu, Kannala, & Borji, 2019) performed several tests including a test regarding vanishing point in predicting gaze in a driving game, which is relevant for this study. Furthermore, the task-specific features (such as vanishing point) are only predictive while the subject is performing the task (a problem addressed by e.g. (Peters & Itti, 2007), using scene gist to estimate task). At this point, we are in danger of attempting to build a full cognitive model of all behavior (which would of course be necessary to fully model gaze behavior, which is part and parcel of the overall behavior of a subject). Thus, we first address the simpler issue of how bottom-up circuits influence gaze during an arbitrary (real-world) task. Future work will address the isolation of these bottom-up factors from other parallel and interacting factors such as task, subject experience, or memory, which likely have a much stronger effect on gaze in most situations but which will require more sophisticated models and experiments.

In this research, we apply a simple conventional model of bottom-up visual attention (the IK saliency model) and find that saliency does predict gaze better than chance even under top-down task conditions such as during driving. However, gaze is better predicted by saliency under conditions with fewer top-down constraints (riding in passenger seat). Finally, surprisingly, prediction performance is better at night, possibly due to the higher contrast of artificial lighting in the dark. Indeed, the purpose of (safety) lights at night is largely to draw attention to important things. Thus, this result may reflect a higher correspondence/correlation between important things and visually salient (lit up) things at night.

Pre-processing of visual stimuli to project them into retinotopic space slightly improved prediction performance. This is likely due to this method taking advantage of subjects' VOR and OKR to stabilize the visual image on the retina. This cancels out distracting perceived motion and blur due to movement of the subject or the outside world. We were surprised that this did not improve performance better. In hindsight, this is unsurprising given that human adults (unlike e.g. cats) tend to keep their heads relatively stable and upright even under dynamic natural conditions (Einhäuser et al., 2009, 2007; Holt, Ratcliffe, & Jeng, 1999).

Methods

Behavioral Experiments

Subjects drove (or were driven in) a consumer automobile (Mazda CX5) on a fixed public course in Hiroshima, Japan. The course required roughly 15 minutes per loop. Four employees of Mazda Motor Co. voluntarily participated in the experiments (Tab. 1), all being experienced drivers. The experiments were approved by the ethics and safety board at Mazda Motors Corporation. Each subject completed one trial under each experimental condition. Each trial included two back-to-back loops around the course. Trials took place between 2PM and 5PM ("day", which sometimes included dusk) or between 6PM and 9PM ("night") in late February (22-24) in Hiroshima, Japan. All subjects completed one trial for each task condition (driver versus passenger) and each time condition (day versus night), i.e. a total of four trials. However, data collection was only possible for two subjects at night. Weather conditions were clear or cloudy for all experiments. Subjects memorized the course before each trial, and were intimitely familiar with the area containing the course and the roads comprising the course.

Table 1: Subject properties

Subject	Age	Sex
A	30s	M
B	40s	M
C	50s	F
D	60s	M

Equipment and Analysis

Subjects wore a Tobii Pro Glasses 2 (TG2)[1] wearable eye tracker, secured via a safety band around the back of the head.

[1]Tobii AB, Stockholm, Sweden, https://www.tobii.com/products/discontinued/tobii-pro-glasses-2

The TG2 connected wirelessly to a laptop computer running eyerevealer[2] to stream and save video (h264 1920x1080 at 25 Hz), eye (100 Hz), and inertial (IMU – 100 Hz) data.

The streamed TG2 data (stored as MPEG-TS video and JSON files) was resampled and eye and IMU data synchronized with head-camera video. Gaze samples were converted from 3D Euclidian vectors provided by TG2 into visual angles (Euler angles: yaw and pitch).

Gaze Preprocessing/Saccade Detection

Gaze signals were smoothed and interpolated using a median filter and Savitzky–Golay filter (0.039 seconds width) according to the preprocessing steps of the remodnav model (Dar, Wagner, & Hanke, 2021). We furthermore reimplemented the eye movement and event detection algorithms, but found TG2's gaze sample rate of 100 Hz to be too low/noisy for effective saccade detection. This may be especially true due to the large number of small eye movements under the driving conditions. As such, all further analyses in this paper use instantaneous gaze position, rather than saccade endpoints. This has the advantage of not masking situations where gaze continues at a location due to it having high saliency, whether it does so due to a local (micro)saccade, or simply a fixation of increased duration.

Isometric Visual Angle Coordinates

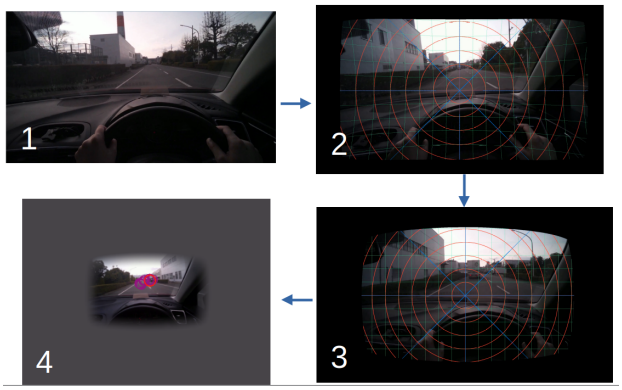

Figure 1: Conversion from raw video to isometric visual angle coordinates. 1) Raw video image. 2) Lens distortion removed (pinhole camera coordinates). 3) Isometric visual angle pixel coordinates. 4) Embedded in mean luminance background and reduced contrast of interface.

Raw head camera videos were transformed so that pixel coordinates correspond to (head-centric) isometric visual angles (Fig 1).

Head-camera video was undistorted based on lens and camera intrinsic parameters to represent a theoretical pinhole camera's image. We then constructed an image representing isometric visual angles by sampling pixels from the pinhole camera image on a regularly spaced grid of visual angles (Euler yaw and pitch angles) projecting outwards from the pinhole. We used a grid of spacing 0.010 degrees of visual angle (dva).

After the preprocessing step described above, the isometric head-camera images represent visual angle in head coordinates, with the center of the image corresponding to straight forward from the head. Head-coordinate images were then embedded in a mean-luminance gray background ($(r+g+b)/3$) using a mask specifying isometric visual angles which hit the image plane and thus contain useful visual information. We next reduced the contrast of image pixels near the interface between the contentful embedded image and the gray background based on their distance to the edge using a Gaussian taper (standard deviation 2.0 dva). A pixel on the image/background interface will have a maximum contrast of zero (100% gray background, 0% image data) and a pixel three standard deviations (6.0 dva) from the image/background interface will have a maximum contrast of 99.7%.

Conversion to Retinotopic Coordinates

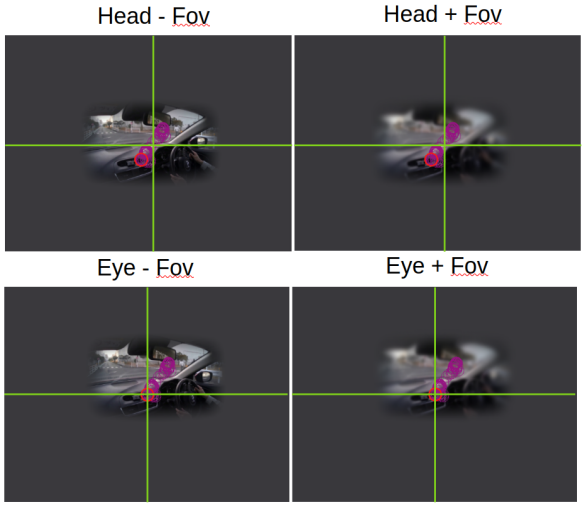

Figure 2: Head- and eye-centered isometric visual angle coordinate images, with (+fov) or without (-fov) foveation. The dark red circle represents the present gaze position in the image.

In a retinotopic (eye-center) image, the center of the image corresponds to the current gaze position. We produced retinotopic (eye-centered) images by counter-rotating isotropic visual angle images based on the subject's gaze-in-head position at the time the video image was captured (Figure 2, bottom). Specifically, pixels in each head-space isotropic visual angle video frame were translated in a direction opposite the average angle of the gaze-in-head for all gaze samples during that video frame, with missing values filled with the same default mean gray luminance.

[2] https://github.com/flyingfalling/eyerevealer

We furthermore applied a foveation model to the visual images. A foveation model resamples the image to mimic the reduced acuity observed in subjects as a function of visual eccentricity (and, in complex models, angle). Parts of the image further away from the current gaze location are blurred more. We constructed foveated versions of both the head-space and the eye-space (isometric) videos (Figure 2, right) to dissociate the effect of foveation from that of eye-centering. We specifically implement the foveation method of (Perry & Geisler, 2002), with the blur wavelength (degrees per cycle = dpc) rising linearly from the center of gaze in every direction with a constant slope of 0.020 dva blur per additional dva eccentricity, and an intercept of 40.0 cpd.

For video frames without corresponding gaze samples (due to blinks, etc.), the entire output frame is a constant gray with the mean luminance of the (unshown) video frame.

Saliency Methods

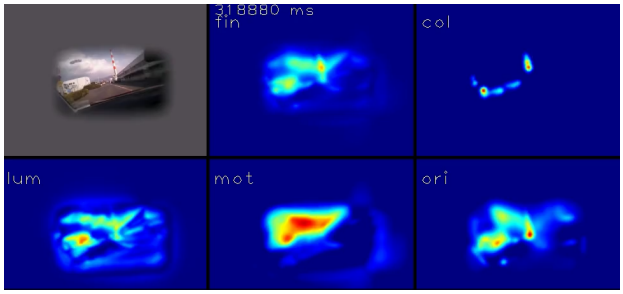

Figure 3: Top-left: Input image (isometric visual angle coordinates). "fin": combined saliency, "ori": orientation feature channel, "col": color, "mot": motion, "lum": luminance.

Visual saliency was computed via an implementation of the Itti-Koch saliency map model (IK) (Itti & Koch, 2000) implemented in salmap_rv [3]. Attention models and saliency maps often operate in image (pixel) space, with the assumption that subjects will dynamically modify the parameters of the saliency maps in their brain to adapt to the relative properties and size of the content. While this has yet to be proved, subjects are known to dynamically adapt other aspects of behavior (saccade amplitude) based on stimulus and task properties (Rothkegel, Schütt, Trukenbrod, Wichmann, & Engbert, 2019). This assumption is not problematic when subjects view clearly-delineated visual stimuli (images, videos) on a computer screen, but does not hold up when one considers that a subject moving around in the natural world will have the full visual field as their "content", and the size of the region on which they focus may depend upon the task or other considerations. We computed the saliency map model with luminance, orientation (4 angles), color (combined RG/BY), and visual motion (4 directions, starting at 2.0 dva/sec velocity at level 0), with the first level having frequency of 1.2 cpd, and every subsequent level half that (4 centers starting at level 1, 2 surround level per center, offset starting at 3 levels from center). Map competition was accomplished via iterative Difference-of-Gaussian winner-take-all method (Itti & Koch, 2000) (center gaussian sigma 1.0 dva, excite weight 0.5, surround sigma 12.0 dva, inhib weight 1.5, 4 iterations, constant inhibition 0.01).

[3] https://github.com/flyingfalling/salmap_rv

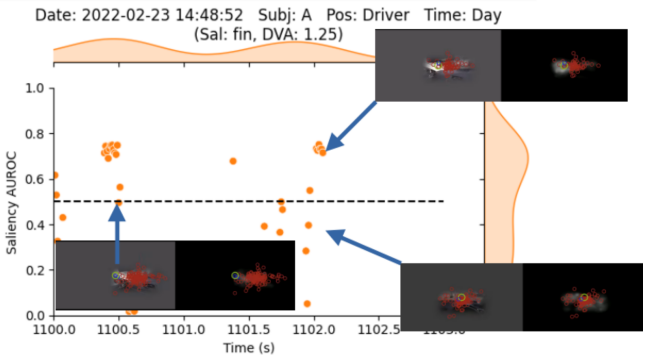

Figure 4: Visualization of the method to compute the percentile (i.e. AUROC) for each gaze sample. Each of the three inserts shows a visual input (left) and saliency map (right). This example visual input is eye-centered, and the current target of gaze is shown by the large GREEN circle at the center of each image. Random draws from this subject's other looking locations in head-coordinates are shown as RED circles. The gaze position of the NEXT look (100 ms later) is shown in BLUE. Saliency AUROC is the percentile (scaled to [0,1]) of the saliency values within the BLUE circle within in the set of saliency values within the RED circles.

The IK saliency map computes regions of an image which are locally conspicuous based on difference between features present at fine-grained central regions versus coarse-grained surround regions at multiple spatial scales. Separate maps are computed for each feature channel aspect (each orientation angle, each motion direction, each color opponency channel) at multiple center-surround spatial scales, then combined via spatial competition into a map representing the conspicuous regions in that feature channel (i.e. most conspicuous areas of motion, orientation, color, regardless of the direction, angle, or opponency that caused it) (Fig. 3, "mot", "lum", "col", "ori"). Finally, the feature channels are combined again via spatial competition into a final map (Fig. 3, "fin") representing the regions of the image which differ most from their surroundings in general.

We apply this saliency map model to every frame of the isometric visual angle coordinate space videos produced for every foveation/centering condition (head+fov, eye-fov, etc.). The results are output as 1.0 dva/pixel images. We then determine whether the image locations where a subject looked have have higher saliency values than locations where the subject did not look (drawn from a null model which is the subject's prior distribution of looking locations across all trials, see Fig. 4). This is the "shuffled AUC (sAUC)" method

(Bylinskii et al., 2018). Note that to account for visual processing and motor coding and execution delay, we sample the visual saliency of the corresponding retinotopic location from a timepoint $\delta_t = 100$ milliseconds *before* the timepoint of the gaze sample. This is especially important for the eye-centered condition, since the video location specified by the current gaze location will change depending on how the subject moves their eyes. 100 ms is a safe value which takes into account the maximum time for executing a saccade (20-50 ms) as well as other factors such as visual propogation (50-70 ms to superior colliculus or primary visual cortex, where saliency is thought to be computed/represented (White, Berg, et al., 2017; White, Kan, Levy, Itti, & Munoz, 2017; Veale, Hafed, & Yoshida, 2017)).

In reality, we determine the percentile of the gazed location's saliency value within the null model of all gazed location's saliency values, which corresponds to the AUROC when one has only one true positive value. The mean of all such percentiles corresponds to the overall AUROC of all gaze positions (i.e. true positives). An AUROC of 0.5 means that the saliency map model predicts gaze no better than the null model (i.e. the subject's prior looking distribution). An AUROC over 0.5 indicates that subjects looks are on average better predicted by saliency than the null model. And AUROC less than 0.5 mean that the subject is intentionally looking away from salient targets.

Results

Behavioral Phenomenology

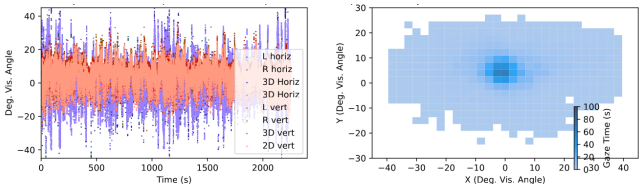

Figure 5: Left: Gaze position over the time course of a single trial represented separately as horizontal and vertical angle in head. Right: Density of head-centric gaze positions – a tendency for the eyes to look straight ahead is clear.

Subjects tend to keep their eyes focused within the central 15°, although some looks occur up to about 40°horizontally (the limit of our measurement) and 25°vertically (Fig 5). Previous reports have established that large gaze shifts are often accomplished by a combined head-eye (and sometimes trunk/torso) movement along with counter-rotation of the eye in the head, and that humans prefer to keep the eyes in a comfortable position relative to the head (Radau, Tweed, & Vilis, 1994; Crawford & Vilis, 1991; Land, 2006). Our observed gaze distributions lie comfortably within this region.

Saliency Prediction of Gaze

While the AUROC of saliency at predicting gaze varies from moment-to-moment (sample-to-sample), the average over the trial is greater than chance (Fig. 6 for example trial). Preprocessing conditions (centering and foveation) have a weak effect on saliency prediction performance, with the largest observed improvement over the default head-fov being eye+fov, which improves on AUROC by 0.014 in night driving conditions (Fig. 7). Taking the average improvement over all conditions, we find that applying the eye+fov preprocessing causes significantly improved AUROC over head-fov condition (right-tailed t-test(df=3, t=3.1779), p = 0.025087).

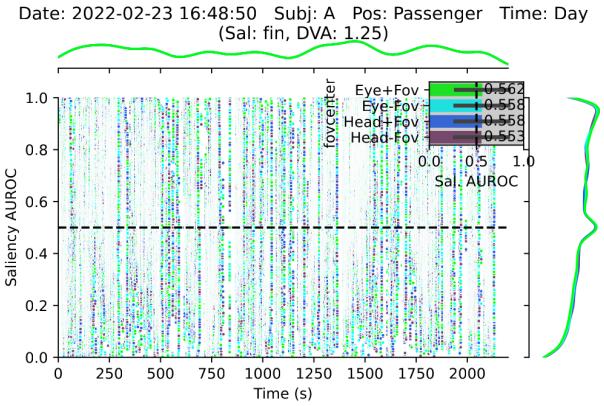

Figure 6: AUROC predicted by saliency model for each of the centering/foveation conditions for each time point of an example trial (Day, Passenger). Saliency sampled 1.25 dva Gaussian around gaze location. Right: marginal distribution of AUROC of all samples in this trial. Note it is similar among the different foveation/centering conditions (mean and standard deviation shown in inset for each condition).

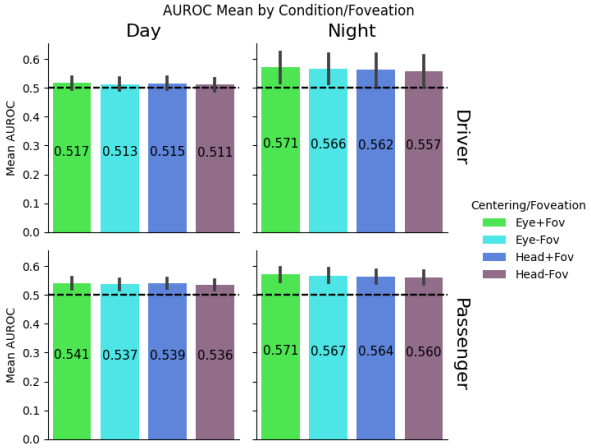

Figure 7: Mean and variance of AUROC among subjects for each task condition, driving condition, and foveationg/centering condition.

Similarly, environmental conditions (day/night) and task condition (driver/passenger) have a significant effect on mean

AUROC (Fig. 7). In general, passenger gaze is better predicted by saliency than driver gaze (especially during the day, +0.024). Saliency better predicts gaze during the night (max of +0.054 AUROC for driver night over day).

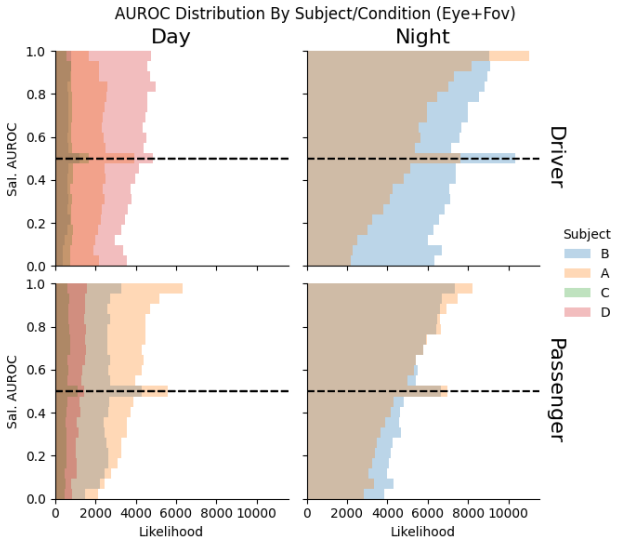

Figure 8: Marginal distributions of per-sample percentiles (AUROC).

Digging into the cause of this, we see the marginal distribution of the percentiles of individual gaze samples is highly skewed towards extreme values (1.0) in the night (Fig. 8, right two plots), whereas in the day it is more balanced (left two plots). The large number of high-percentile samples indicates that a highly salient thing drew gaze, and that there were not many other salient things in the visual field to look at (at least around where the subject usually looked).

Note that the large peak at exactly 0.5 AUROC in all marginal plots corresponds to gaze positions predicted by an all-gray visual stimulus (due to being in blink or losing tracking 100 milliseconds before the current gaze sample). These samples could be removed to improve mean AUROC. This is analogous to a situation where a subject looks to a spatial position for which we have no visual data (lying outside the camera field of view, e.g. looking to a target at 50 degrees eccentricity). We do not exclude these large looks even though the target location has artificially reduced saliency due to our not having access to the visual stimulation available to the subject at the time of choosing to look there.

While task and environmental conditions cause different predictivity of saliency, there is also variance between subjects even within the same task and environmental conditions (Fig. 9). For example, subject A gaze was predicted better by saliency than subject B while driving at night (+0.075 AUROC).

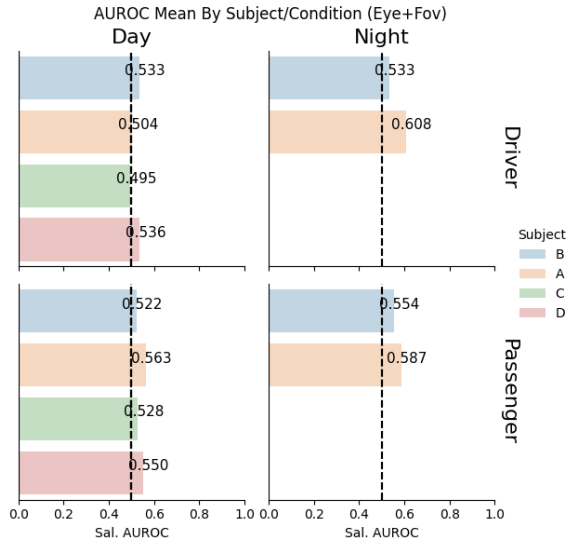

Figure 9: AUROC of best model (eye+fov) shown separately by subject, task, and driving condition.

Discussion

The ability of a bottom-up visual attention model (IK saliency map) to predict gaze is maintained even under real-world natural task conditions such as operating or being a passenger in a vehicle, albeit the predictivity of the model changes depending on task conditions. Predictive ability is on par with previously reported values for IK saliency prediction in the laboratory (from 0.53 to 0.65 in human adults depending on the visual stimulus and report) (Bylinskii et al., 2018; Chen et al., 2021).

Recently, more performant models use information theoretic (Bruce & Tsotsos, 2007) or (pre-learned) statistical principles to determine which features draw gaze (SUN (Zhang, Tong, Marks, Shan, & Cottrell, 2008)), or learn arbitrary function approximators (convolutional neural networks such as DeepGaze or SalGan (Pan et al., 2017)). Such models have achieved high performances up to 0.77 AUROC (human interobserver models achieve 0.81, i.e. this is the theoretical maximum for a model which does not adapt based on behavior and which represents the average subject's behavior, rather than a specific subject). These models are trained on human adult behavior and thus encode not only bottom-up visual features, but also social, cultural, and ecological norms, as well as learned experiences of subjects about the world (e.g. gravity, statistical regularities in modern architecture) (Hayes & Henderson, 2021). The IK model is simple and mimics basic visual processes in early visual areas (White, Berg, et al., 2017). We have shown that it can be extended to real-world data from wearable eye trackers in naturalistic conditions, rather than stimuli presented in the laboratory. We hope that the study of embodied behavior will continue to expand and that models will be evaluated under more realistic conditions such as those presented here.

Acknowledgments

This research was partially supported by a collaborative research grant from Mazda Motor Corporation to TI and RV. Software packages were developed in part with support from KAKENHI grant 21K15609 to RV.

References

Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, *37*(6), 1453–1467.

Borji, A., & Itti, L. (2014). Defending yarbus: Eye movements reveal observers' task. *Journal of vision*, *14*(3), 29–29.

Bruce, N., & Tsotsos, J. (2007). Attention based on information maximization. *Journal of Vision*, *7*(9), 950–950.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, *41*(3), 740–757.

Chen, C.-Y., Matrov, D., Veale, R., Onoe, H., Yoshida, M., Miura, K., & Isa, T. (2021). Properties of visually guided saccadic behavior and bottom-up attention in marmoset, macaque, and human. *Journal of Neurophysiology*, *125*(2), 437–457.

Crawford, J., & Vilis, T. (1991). Axes of eye rotation and listing's law during rotations of the head. *Journal of neurophysiology*, *65*(3), 407–423.

Dar, A. H., Wagner, A. S., & Hanke, M. (2021). Remodnav: robust eye-movement classification for dynamic stimulation. *Behavior research methods*, *53*(1), 399–414.

Einhäuser, W., Moeller, G. U., Schumann, F., Conradt, J., Vockeroth, J., Bartl, K., ... König, P. (2009). Eye–head coordination during free exploration in human and cat. *Annals of the New York Academy of Sciences*, *1164*(1), 353–366.

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., & König, P. (2007). Human eye-head coordination in natural exploration. *Network: Computation in Neural Systems*, *18*(3), 267–297.

Hayes, T. R., & Henderson, J. M. (2021). Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific reports*, *11*(1), 18434.

Holt, K. G., Ratcliffe, R., & Jeng, S.-F. (1999). Head stability in walking in children with cerebral palsy and in children and adults without neurological impairment. *Physical therapy*, *79*(12), 1153–1162.

Huang, Y., Cai, M., Li, Z., Lu, F., & Sato, Y. (2020). Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, *29*, 7795–7806.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.

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

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of vision*, *14*(3), 14–14.

Kotseruba, I., & Tsotsos, J. K. (2022). Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE Transactions on Intelligent Transportation Systems*, *23*(11), 19907-19928. doi: 10.1109/TITS.2022.3186613

Kübler, T. C., Kasneci, E., Rosenstiel, W., Heister, M., Aehling, K., Nagel, K., ... Papageorgiou, E. (2015). Driving with glaucoma: task performance and gaze movements. *Optometry and Vision Science*, *92*(11), 1037–1046.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, *25*(3), 296–324.

Land, M. F. (2009). Vision, eye movements, and natural behavior. *Visual neuroscience*, *26*(1), 51–62.

Land, M. F. (2015). Eye movements of vertebrates and their relation to eye form and function. *Journal of Comparative Physiology A*, *201*(2), 195–214.

Lappi, O. (2016). Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference. *Neuroscience & Biobehavioral Reviews*, *69*, 49–68.

Lappi, O. (2022). Gaze strategies in driving–an ecological approach. *Frontiers in psychology*, *13*.

Mao, D. (2023). Neural correlates of spatial navigation in primate hippocampus. *Neuroscience Bulletin*, *39*(2), 315–327.

Mao, D., Avila, E., Caziot, B., Laurens, J., Dickman, J. D., & Angelaki, D. E. (2021). Spatial modulation of hippocampal activity in freely moving macaques. *Neuron*, *109*(21), 3521–3534.

Ottes, F. P., Van Gisbergen, J. A., & Eggermont, J. J. (1986). Visuomotor fields of the superior colliculus: a quantitative model. *Vision research*, *26*(6), 857–873.

Palazzi, A., Abati, D., Solera, F., Cucchiara, R., et al. (2018). Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, *41*(7), 1720–1733.

Pan, J., Canton, C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., & Giro-i Nieto, X. a. (2017, January). Salgan: Visual saliency prediction with generative adversarial networks. In *arxiv*.

Perry, J. S., & Geisler, W. S. (2002). Gaze-contingent real-time simulation of arbitrary visual fields. In *Human vision and electronic imaging vii* (Vol. 4662, pp. 57–69).

Peters, R. J., & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 ieee conference on computer vision and pattern recognition* (pp. 1–8).

Radau, P., Tweed, D., & Vilis, T. (1994). Three-dimensional

eye, head, and chest orientations after large gaze shifts and the underlying neural strategies. *Journal of Neurophysiology*, *72*(6), 2840–2852.

Rothkegel, L. O., Schütt, H. H., Trukenbrod, H. A., Wichmann, F. A., & Engbert, R. (2019). Searchers adjust their eye-movement dynamics to target characteristics in natural scenes. *Scientific reports*, *9*(1), 1–12.

Schumann, F., Einhäuser, W., Vockeroth, J., Bartl, K., Schneider, E., & König, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of vision*, *8*(14), 12–12.

Takahashi, M., & Veale, R. (2023). Pathways for naturalistic looking behavior in primate I: Behavioral characteristics and brainstem circuits. *Neuroscience*, *532*, 133-163.

Tavakoli, H. R., Rahtu, E., Kannala, J., & Borji, A. (2019). Digging deeper into egocentric gaze prediction. In *2019 ieee winter conference on applications of computer vision (wacv)* (pp. 273–282).

Veale, R., Hafed, Z. M., & Yoshida, M. (2017). How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160113.

White, B. J., Berg, D. J., Kan, J. Y., Marino, R. A., Itti, L., & Munoz, D. P. (2017). Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nature communications*, *8*(1), 1–9.

White, B. J., Kan, J. Y., Levy, R., Itti, L., & Munoz, D. P. (2017). Superior colliculus encodes visual saliency before the primary visual cortex. *Proceedings of the National Academy of Sciences*, *114*(35), 9451–9456.

Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., & Hiraki, K. (2012). Attention prediction in egocentric video using motion and visual saliency. In *Advances in image and video technology: 5th pacific rim symposium, psivt 2011, gwangju, south korea, november 20-23, 2011, proceedings, part i 5* (pp. 277–288).

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), 32–32.