UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Stimulus size influences gaze targets during free viewing of natural video

Permalink
https://escholarship.org/uc/item/6416016s

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Wakai, Daiki
Isa, Tadashi
Veale, Richard E

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Stimulus size influences gaze targets during free viewing of natural video

Daiki Wakai (wakai.daiki.28f@st.kyoto-u.ac.jp)
School of Medicine, Kyoto University

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

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

Abstract

While modern computational models accurately predict free viewing scanpaths via learned deep neural networks, the investigation of the brain's implementation of the modeled behavior has been thus far limited to primarily bottom-up visual features. One key question in investigating the early visual system's implementation of bottom-up visual saliency is how (and whether) it dynamically adapts its sensitivity to features of different spatial scales. This paper provides a simple test of whether identical stimuli presented at different spatial scales produce different gaze behavior. We asked subjects (n=12) to freely view video stimuli twice each session over two sessions In one session (intervention) the visual scale changed from large (25 degrees of visual angle) to small (10 degrees) between viewings. In the other session (control) the video size was the same. Gaze was more strongly correlated between viewings of the same video size (r=0.265) than different sizes (r=0.231), independent of whether there was a long ($>$ 24 hours) or short ($<$ 10 min) delay between viewings, implying that memory effects are not a strong factor. Although low, these within-subject correlations are higher than the correlation of gaze between different subjects viewing identical videos (r=0.195).

Keywords: Visual Saliency; Bottom-up Visual Attention; Saliency Map; Visual Cortex; Eye Tracking

Introduction

Animals, including humans, move their eyes constantly to gather information about the world (Takahashi & Veale, 2023; Veale & Takahashi, 2024). Eye movements enable the stabilization of visual input against self-motion and external motion. Eye movements also bring interesting targets into the high-acuity foveal region of the retina, enabling better discrimination, change detection, and visual feature analysis (Land, 2015, 2009). In human, the neural circuits underpinning oculomotor control are organized such that foveating eye movements have special status such as increased speed and accuracy over non-foveating saccades after macular degeneration (Whittaker, Cummings, & Swieson, 1991) or e.g. the quick phases of optokinetic nystagmus (Garbutt et al., 2003). Furthermore, the foveal visual field connects to specialized neural circuits through rostral superior colliculus (SC) and raphe nucleus to suppress eye movements away from a fixated visual target when correct conditions are met (Takahashi, Sugiuchi, Izawa, & Shinoda, 2005). This organization of the visuo-oculomotor system suggests that the allocation of gaze, especially foveating gaze, is an important function of the brain, and presumably fulfills important functions for intelligent behavior. Computational models following the original "saliency map" model (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000) have iteratively improved on performance in predicting human and animal gaze behavior, approaching the problem from a variety of theoretical perspectives. The simplest bottom-up models such as the original saliency map (Itti & Koch, 2000) use fixed-width, hand-crafted visual filters (responding to e.g. luminance differences, oriented features, color differences, or local motion, based on the qualities of neural tunings in early visual cortex (Hubel & Wiesel, 1968), which have since been shown to correlate with saliency (White, Berg, et al., 2017; White, Kan, Levy, Itti, & Munoz, 2017)). More recent models use an end-to-end approach, training deep neural networks to approximate the function mapping visual input to gaze distribution (Huang, Shen, Boix, & Zhao, 2015; Kümmerer, Bethge, & Wallis, 2022; Borji, 2019; Liu & Han, 2018; Pan et al., 2017).

These models achieve admirable results predicting gaze on benchmarks such as the MIT or CAT2000 datasets (Borji & Itti, 2015; Bylinskii et al., n.d.). Saliency models have been used in widespread research to describe the behavior of non-human primates (Shepherd, Steckenfinger, Hasson, & Ghazanfar, 2010) or the effect of cortical blindness (Yoshida et al., 2012), or to identify differences in brain activity in schizophrenia (Nardo, Console, Reverberi, & Macaluso, 2016). All studies applied similarly parameterized saliency map models, despite the visual stimuli varying widely in size, from 61.6 degrees of visual angle (dva) (Yoshida et al., 2012), to 19 dva (Nardo et al., 2016), to 17 dva (Shepherd et al., 2010). Benchmarks are collected at roughly 50.5 dva width (Borji & Itti, 2015). The parameters of most saliency map models are defined in terms of "pixels" and not in terms of "visual angle", and thus the spatial scale of a model is not determined except when applied to a particular visual stimulus presentation size (e.g. (Yoshida et al., 2012) specify the spatial scales of the saliency map used in cycles/degree visual angle; (Chen et al., 2021) explicitly specify visual angle scales). This assumption is not usually made explicit (at least within the model itself), leading to misunderstandings and difficulty in reproducing results using stimuli from the same database but at different sizes, or using differently-parameterized saliency map models. For example, the excellent deepgaze and evaluation implementation[1] mentions

[1] https://github.com/matthias-k/DeepGaze

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

briefly only near the end of the README:

> Please note that all DeepGaze models have been trained on the MIT1003 dataset which has a resolution of 35 pixels per degree of visual angle and an image size of mostly 1024 pixel in the longer side. Depending how your images have been presented, you might have to downscale or upscale them before passing them to the DeepGaze models.

In contrast, salmaprv[2] explicitly requires specification of the dva/pixel of input perceptions, and also requires specification of model parameters in terms of dva.

The convention of dynamically resizing the visual stimuli to the model-space can be interpreted as a theoretical assumption that the brain of subjects is capable of dynamic modulation of its input such that feature tuning and interactions between visual features depend on the task and visual stimulus. It is known that humans are capable of dynamically adjusting the visuo-oculomotor system based on the particular demands of a task or expected stimulus (Rothkegel, Schütt, Trukenbrod, Wichmann, & Engbert, 2019; Kadosh & Bonneh, 2022). However, it is not clear whether the "bottom-up" saliency map (presumed to be an unconscious, automatic filter and competition of visual input to draw gaze to conspicuous targets) is also dynamically adjusted. For example, irrelevant scales may be suppressed dynamically based on context such as video size.

In this paper, we address this question by observing gaze correlation while subjects watch the same video at the same or different size. Furthermore, to dissociate the effect that memory (experience) has on gaze targets to a video, we show videos of identical or different sizes to subjects after either a short (minutes) or long (days) delay contrast gaze correlation between these two memory conditions as well.

Hypothesis: Bottom-up attention has natural and fixed spatial scales

The motivation for these experiments is to differentiate two hypotheses (Fig. 1):

1. "Dynamic saliency": Human subjects dynamically adapt spatial attentional mechanisms depending on the scale of the stimulus or context such as screen size. *Gaze to videos of the same versus different sizes should be equally correlated.*

2. "Static saliency": Human subjects have static, finite attentional mechanisms based on evolutionary and/or learned spatial scales which can not be dynamically adapted. *Correlation to videos of the same size should be higher than to videos of different sizes.*

The experiments proposed in this study address this in the simplest way. However, there are several confounds:

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

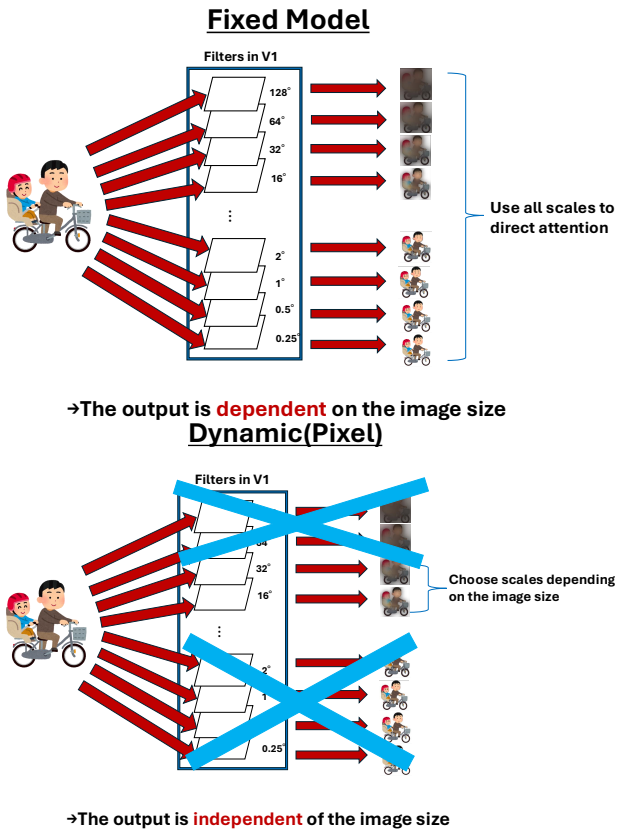

Figure 1: Model hypotheses: Fixed versus Dynamic. Here, primary visual cortex is used as an example for a brain area that might implement the relevant neural computations.

Behavioral Confounds

1. Motor confound: animals have general (least effort) preferences for movements, including gaze movements. These preferences are derived from personal/idiosyncratic preference, experience, subject morphology, and species-specific properties. The preference for maintaining eye-in-head position near the neutral head-forward position is well-documented (Freedman & Sparks, 1997; Radau, Tweed, & Vilis, 1994; Misslisch, Tweed, Fetter, & Vilis, 1994; Guitton & Volle, 1987). Thus, even if the "dynamic saliency" model obtains, gaze will differ between small and large videos due to competition between the "perceptual" saliency pulling gaze to identical targets in both videos and "motor" preferences which will dampen some more eccentric gaze positions. However, recent research has clarified that motor biases are not the primary vehicle for biasing gaze (Tseng, Carmi, Cameron, Munoz, & Itti, 2009), but it can become more relevant under when effort is modulated (Koevoet et al., 2025).

2. Memory confound: A subject viewing a video repeatedly will have a memory of the first time. Gaze patterns could change, either due to prediction of contents (looking to a location before something appears there) or due to the pref-

erence to look to novel parts of the video. This may interact with video contents, as some videos may be naturally more "memorable" than others (spatial scales and layout of images has been shown to modulate both memory and time perception (Ma, Cameron, & Wiener, 2024)).

Experimental/Physical Confounds

1. Video-size versus feature-size confound: Videos were scaled up or down without controlling for the Nyquist frequency of the smallest features visible in the videos. The original video contents were 800x600 pixels, and the small videos were shown at 417x312 pixels, the large at 1042x781 pixels.

2. Pixel-size confound: since we use an identical monitor to show both large and small videos, the spatial scale (Nyquist frequency) of the smallest discriminable features is fixed, effectively high-pass filtering (dynamic) spatial scales when the video is shown very small.

3. Net-luminance confound: since larger videos will by definition have more content, and the surrounding context (gray border around videos) is not modulated to negate this, net luminance differs between conditions.

Methods

Subjects viewed 60 short videos while we measured gaze positions. Subjects viewed each video a total of four times, at a subjective visual width of either 25 degrees of visual angle (dva), or 10 dva. Videos were resized using OpenCV's (Bradski, 2000) "resize" (default interpolation method).

Materials: Video stimuli

We drew video clips from a video clip database selected using pyviddbcreator[3]. Videos in the database produced by automatically segmented 10-second clips excluding scene cuts (via PySceneDetect[4]), followed by manual selection of a database of 200 clips to contain a range of contents, colors, and themes (only 60 are used in the present study). The clips are luminance-scaled via tone-mapping (Mantiuk, Daly, & Kerofsky, 2008) to prevent excessive pupil dilation changes. All clips have a 4x3 aspect ratio at 800/600 width/height pixel resolution in the database. The videos are divided into "sets" of 30 videos each (5 minutes when shown back-to-back). In these experiments we used two of the 30-clip sets of the videos (sets C and D, examples Fig. 2).

Video stimuli are encoded at 30 frames per second (fps) and displayed at this rate. Stimulus display is accomplished via psychopy (Peirce, 2009), using the "freeviewing" module of the peyeutils library[5]. Stimuli were displayed centered on the monitor surrounded by 50% grey uniform luminance background.

[3] https://github.com/flyingfalling/pyviddbcreator
[4] https://github.com/Breakthrough/PySceneDetect
[5] https://github.com/flyingfalling/peyeutils

Figure 2: Example scenes from the video clips.

Methods: Subjects

14 subjects were recruited from Kyoto University and surroundings within a 2-month period. 11 subjects (6 female) completed both experimental conditions and 1 subject completed one condition, and were used for analysis. Mean age was 24.5 years (std: 6.85).

Methods: Behavioral Protocol

Subjects sat at a desk and stabilized their head using a chinrest (SR Research, Canada) with a loose head strap. A computer monitor (Toshiba 15-inch LCD monitor) mounted on a pneumatic arm was fixed so that the top half of the monitor was level with the subject's eyes at a distance of 0.7 m. An Eyelink 1000+ "desktop mount" using the 35 mm lens and 890 nm illuminator was fixed to the table below and in front of the monitor. Gaze was tracked binocularly at 1000 Hz. Room lights were dimmed during experiments. Calibration was accomplished the EYELINK's basic 9-point rectangular calibration with 0.5 dva white annuli with black centers.

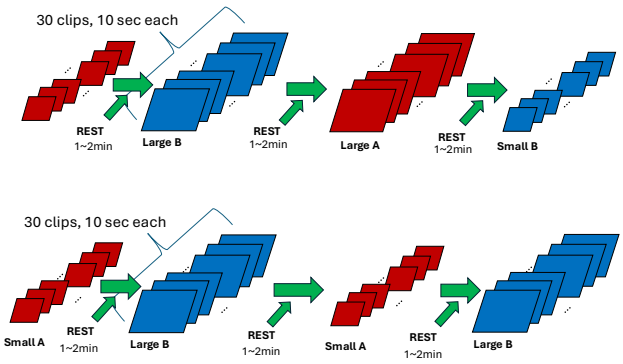

Figure 3: Experimental protocol for intervention (top) and control (bottom) (A and B are arbitrary names for two different sets of 30 video clips)

Fig. 3 summarizes the session/block protocol. Each block

comprised of an eye-tracker calibration followed by back-to-back viewing of a set of 30 video clips (from either the C or D sets) in random shuffled order. Subjects were instructed to "Freely watch the video clips", and instructed to blink freely but not to move their head until the block ended. Either the C or D set video clips were shown first and was randomly counter-balanced between subjects and conditions. After viewing the first set of C or D video clips, the subject rested for about 1 minute (depending on fatigue, blocks were subject-initiated). The rest was followed by another eye-tracker calibration and viewing of the other block (D or C).

After viewing both the C and D sets of videos, there followed a longer intermediate (3-5 minute) "distractor" rest period in which experimenter chatted with subjects about daily life, and viewed irrelevant news-related video clips with subjects on youtube. This distracted subjects from remembering the video clips or thinking too deeply about the experiment.

Following the distractor period, subjects viewed both sets of video clips again (randomly shuffled order), following an identical protocol to the pre-distractor blocks. However, videos were shown at either the same (control session) or different (intervention session) size relative to the pre-distractor block. Since subjects viewed each video clip only four times, each video set in the "control" condition was randomly assigned to be either 25 dva or 10 dva size. In other words, subjects viewed each video clip at a given size (either 10 or 25 dva) three times (once in the intervention, and twice in the control), and the other size (25 or 10 dva) once (once in the intervention).

Subjects were invited back for the second session on a different day. Subjects were reimbursed for participation (2000 JPY).

Methods: Eye Tracking

We tracked subject gaze position using an EYELINK 1000+ (SR Research, Canada). In some subjects, either the left or right eye's data was excluded due to calibration errors. For time points where binocular gaze was unavailable, its position was imputed by adding the average offset between the eyes for that video to the available eye. Gaze was calibrated before each block of videos using a 9-point calibration, with a validation step.

Following data collection, gaze samples were converted to CSV format using pyedfread [6], and smoothed via a median filter. Blinks and noisy pupil data were removed using the pupil size method of (Kret & Sjak-Shie, 2018), and saccades were detected and excluded from correlations using a velocity threshold and EYELINK's default event detection. Correlations were computed using numpy's "corrcoef" method independently for horizontal and vertical eye position and the mean of the two is reported. Correlation computation excluded time points during blinks or fast eye movements such as saccades, which were replaced with NaNs before analysis. This exclusion is justified because excepting very specific conditions (Ibbotson & Krekelberg, 2011), the eye position *during* saccades is unrelated to visual content (i.e. the visual content halfway between target A and B is traversed during a saccade, regardless of what is there). While we do not apply it in this paper, one reasonable extension is to low-pass filter gaze position over time and/or space. This would estimate correlations more robustly under the assumption that the precise timing/targets of gaze are noisy between viewings.

[6] https://github.com/s-ccs/pyedfread

Results

Subjects made natural eye movements while viewing videos of all sizes. Example space-time trajectories for a single subject on a single video are shown in Fig. 4.

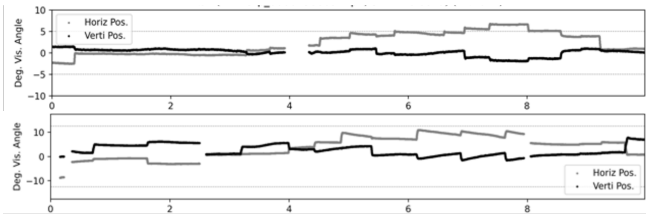

Figure 4: Example traces for the same video viewed by the same subject at different sizes (Top: 10dva, Bottom: 25dva). X-axis is time (sec). Note difference in dva scale (left axis).

Subjects on average looked near the center of the videos. Horizontal deviation of trial mean horizontal eye position (and stddev between trials) for 10 dva videos was 0.00 (stddev: 0.025) dva for intervention condition and 0.06 (0.025) for control condition. For 25 dva videos, mean horizontal position was 0.02 (0.02) (intervention) and 0.025 (0.02) (control) respectively. Vertical gaze mean (stddev) deviation from video center was, respectively, 0.083 (0.03) for 10dva intervention, 0.078 (0.025) for 10dva control, 0.09 (0.02) for 25dva intervention, and 0.12 (0.022) for 25dva control. These deviations of mean gaze from the center are on the scale of 0.1% of video width, indicating that videos were balanced and video size did not significantly affect the neutral gaze position. While in human adults it is usually assumed that subjects behaved as expected (i.e. looked at the videos), this is not always the case with animal or infant studies and this provides a sanity check for our data.

Result: Same/different video sizes

We computed the correlation between the gaze positions of subjects viewing same-sized (control) or differently sized (intervention) videos. We ignore the order in which differently-sized videos were viewed. On first glance, our primary variable of interest (videos being same sized or not) appears to follow the hypothesis of the static model, i.e. differently sized videos should have lower correlation (Fig. 5). Mean correlation coefficient (Pearson's product-moment, r) was 0.265 for control (same size) trials, and 0.231 for intervention (different size) trials. However, several other factors can confound this effect, one being memory/prediction effects caused by

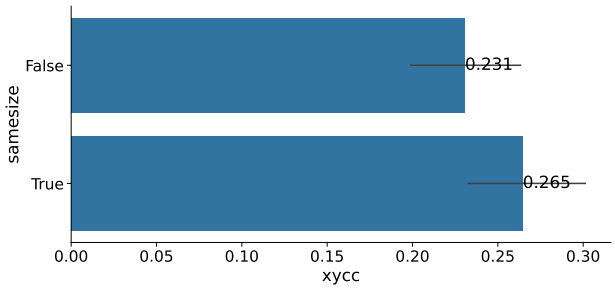

Figure 5: Mean within-subject correlation of gaze position while viewing identical videos at same (10/10 dva or 25/25) or different (10/25 or 25/10) sizes. Errors are 95% CI.

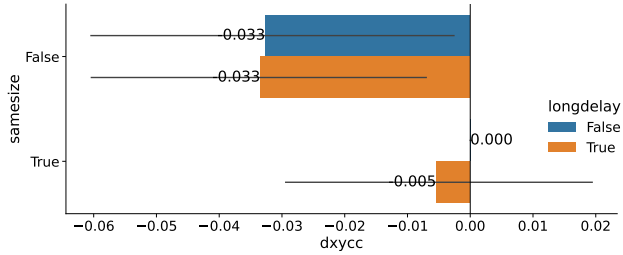

Figure 7: Delta (difference from reference condition of same size/short delay) of mean within-subject correlation of gaze position while viewing identical videos at same (10/10 dva or 25/25) or different (10/25 or 25/10) sizes. Errors are 95% CI.

viewing the same stimulus multiple times. We assume that these effects will be modulated by time, and should specifically be stronger for more recent viewings of a video, although novelty preferences and familiarity/prediction preferences may both be present and result in unexpected effects on the gaze correlation. Unable to easily disentangle this using the present design, we simply compare short ($<$ 10 minutes) versus long ($>$ 24 hours) time delays between viewings as a factor in our analysis. Additional factors such as stimulus memoribility or arousal may have additional interactions with time delay and memory effects, but we ignore them for the present analysis.

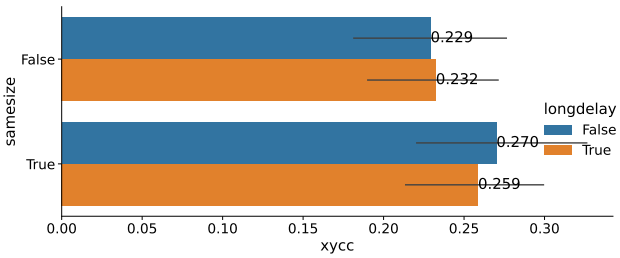

Figure 6: Mean within-subject correlation of gaze position while viewing identical videos at same (10/10 dva or 25/25) or different (10/25 or 25/10) sizes, as a function of time elapsed between viewings being long ($>$ 24 hours) or $<$ 10 minutes.

To test our hypothesis that the delay between viewings has less effect than the video size difference (Fig: 6). Our dependent variable is *xycc*: the mean of the Pearson's correlation of the horizontal and vertical components of gaze. A two-way ANOVA directly predicting xycc given the two independent factors samesize and longdelay reported significant main effect only in samesize (F(df=1,3700)=14, p=0.0001), but not longdelay (F=0.34, p=0.558) or samesize-longdelay interaction (F=1.177, p=0.278). Thus, subject gaze during two viewings of a video had higher correlation if the videos were shown at the same size, regardless of elapsed time between viewings.

Additionally, we tested whether within-video and within-subject differences played a significant role (Fig. 7). In this case, every comparison was locked to a reference condition (specifically the same-size and short-delay condition), effecting a paired (repeated measures) comparison. The same trend was observed under these constraints as well. Specifically, different sized video videos on average having 0.031 lower correlation than same sized conditions. Again, a two-way ANOVA supports this observation, with the only significant factor being samesize (F(df=1,2486)=9.13, p=0.0025). Neither longdelay (F=0.0149, p=0.9) nor the interaction (F=0.026, p=0.87) are likely distributed differently than the null model.

However, subjects did display heterogenous patterns of behavior (Fig 8. Averaging by subject before analysis made the correlations weaken significantly. Again, ANOVA indicated that samesize likely has a stronger predictive contribution to gaze correlation, although with unacceptable risk of type-1 error (samesize F(df=1,42)=1.69, p=0.2; longdelay F=0.022, p=0.88; interaction F=0.03, p=0.86). Observing the distribution of subject behavior, this may be due to subjects exhibiting significantly different behavior from one another (a not uncommon issue in free-viewing tasks). Our previous lumped results' significance may thus derive from a majority of subjects displaying our hypothesized pattern without it being a universal trait.

Results: Time delay does not affect gaze correlation

In order to evaluate the influence of memory on gaze, we further separated the eye movement data in each session into 4 groups: different sized video viewings interrupted by a short interval versus a long interval, and same-sized stimuli interrupted by a short versus long interval. We repeated the analysis separately for the two conditions (Fig. 6). While we did not exclude the possibility that there is an interaction between condition and time delay, subject gaze while watching a previously viewed video does not appear to be significantly altered by whether there was a short ($<$ 10 minutes) or long ($>$ 24 hours) delay. This deviates from our prediction

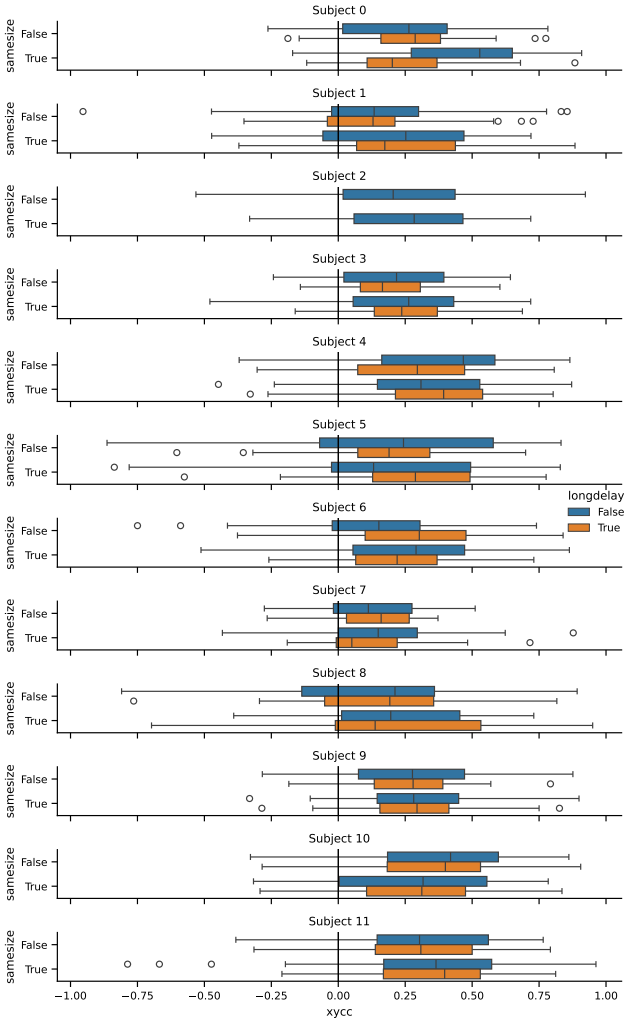

Figure 8: Per-subject distribution of data within long-delay and same-size condition. Note that while most subjects show the main effect of same-sized videos being better correlated (subject 11), some show the opposite pattern (subject 10). Furthermore, subjects show opposite interactions of longdelay, with some crossing conditions one direction (top subject), versus another (e.g. subject 5 versus subject 0).

that longer time delays would result in higher correlations, as short-term memory and recency effects decayed. However, it is possible that multiple competing memory-related processes (predictive gaze versus novelty-seeking) result in a net zero change in correlation for short time delays. For long delays, a corresponding net zero change iscaused by presumably two weaker, yet equally offset, processes. However, separate experiments must be designed to probe this possibility.

Results: Gaze correlation is higher for the same subject than different subjects

We finally investigated the extent to which different subjects shared gaze patterns while watching identical videos. The

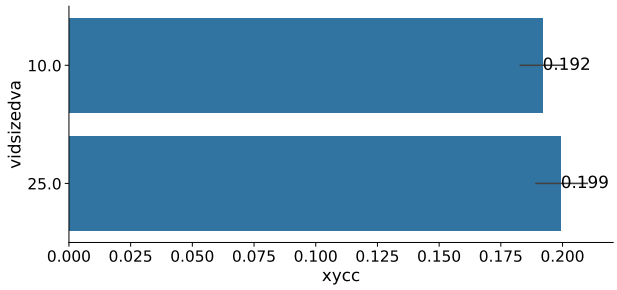

Figure 9: Mean between-subject correlation of gaze position while viewing identical videos at same size.

inter-observer correlation is the highest performance that an "average human" attention model could theoretically achieve when given only visual stimulus input (Fig. 9). We computed the correlations between subjects separately for each video size (25 dva: r=0.199; 10 dva: r=0.192; unpaired t-test t(df=118)=0.52, p=0.60). Thus, as expected, between-subject correlation (mean r=0.195) for identically sized videos is lower than within-subject correlation (r=0.26).

Conclusions and Discussion

We investigated the assumption that human visual attention (as measured by overt gaze) is invariant with regards to the spatial scale of visual stimuli. At extremes, this is obviously not true (a video shown with total width 0.5 dva would be barely visible). However, it has been assumed (implicitly) in multiple reports that the bottom-up attention mechanisms dynamically adapt to video sizes at least between 10 and 60 dva. In this paper, we provided evidence that gaze targets become more uncorrelated even when stimulus size differs between 10 to 25 dva. Thus, it is important for models of visual attention to be explicit about their assumptions regarding spatial scales and the mapping to physical coordinates.

Further experiments are needed to further dissociate the contribution of non-perceptual biases (e.g. eye-in-head centering preferences) on gaze. In addition, a larger range of visual stimulus sizes, and a method of dissociating video size from the spatial scale of its contents will be necessary to better enumerate the multiple processes that are likely influencing gaze in parallel. Finally, statistical estimation of saliency model parameters for each viewing condition will help us separate the different possible models of how the brain implements selective spatial visual attention (i.e. the static versus dynamic model). The ability to model visual attention in real-world situations (e.g. on an egocentric camera with 90 dva field of view (Veale, Murase, Watanabe, & Isa, 2024)) is pushing models to be able to handle a larger dynamic range of stimuli, and the question of whether additional mechanisms are necessary to explain gaze behavior remains an open one.

Acknowledgments

Research funded by JSPS KAKENHI wakate (21K15609) (RV) and AMED Brains/MINDS grant (TI).

References

Borji, A. (2019). Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, *43*(2), 679–700.

Borji, A., & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*. (arXiv preprint arXiv:1505.03581)

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (n.d.). *Mit saliency benchmark.* http://saliency.mit.edu/.

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.

Freedman, E. G., & Sparks, D. L. (1997). Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of neurophysiology*, *77*(5), 2328–2348.

Garbutt, S., Han, Y., Kumar, A. N., Harwood, M., Harris, C. M., & Leigh, R. J. (2003). Vertical optokinetic nystagmus and saccades in normal human subjects. *Investigative ophthalmology & visual science*, *44*(9), 3833–3841.

Guitton, D., & Volle, M. (1987). Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range. *Journal of neurophysiology*, *58*(3), 427–459.

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 ieee international conference on computer vision (iccv)* (p. 262-270). doi: 10.1109/ICCV.2015.38

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Ibbotson, M., & Krekelberg, B. (2011). Visual perception and saccadic eye movements. *Current Opinion in Neurobiology*, *21*(4), 553-558. (Sensory and motor systems) doi: https://doi.org/10.1016/j.conb.2011.05.012

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.

Kadosh, O., & Bonneh, Y. S. (2022). Fixation-related saccadic inhibition in free viewing in response to stimulus saliency. *Scientific Reports*, *12*(1), 1–12.

Koevoet, D., Van Zantwijk, L., Naber, M., Mathôt, S., Van der Stigchel, S., & Strauch, C. (2025). Effort drives saccade selection. *ELife*, *13*, RP97760.

Kret, M. E., & Sjak-Shie, E. E. (2018). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, *51*, 1336 - 1342.

Kümmerer, M., Bethge, M., & Wallis, T. S. A. (2022). Deep gaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of vision*, *22*, 1-27.

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.

Liu, N., & Han, J. (2018). A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE Transactions on Image Processing*, *27*(7), 3264–3274.

Ma, A. C., Cameron, A. D., & Wiener, M. (2024). Memorability shapes perceived time (and vice versa). *Nature Human Behaviour*, 1–13.

Mantiuk, R., Daly, S., & Kerofsky, L. (2008). Display adaptive tone mapping. In *Acm siggraph 2008 papers* (pp. 1–10).

Misslisch, H., Tweed, D., Fetter, M., & Vilis, T. (1994). The influence of gravity on Donders' law for head movements. *Vision research*, *34*(22), 3017–3025.

Nardo, D., Console, P., Reverberi, C., & Macaluso, E. (2016). Competition between visual events modulates the influence of salience during free-viewing of naturalistic videos. *Frontiers in human neuroscience*, *10*, 320.

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.*

Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy. *Frontiers in neuroinformatics*, *2*, 343.

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.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology*, *20*(7), 649–656.

Takahashi, M., Sugiuchi, Y., Izawa, Y., & Shinoda, Y. (2005). Synaptic inputs and their pathways from fixation and saccade zones of the superior colliculus to inhibitory burst neurons and pause neurons. *Annals of the New York Academy of Sciences*, *1039*(1), 209-219.

Takahashi, M., & Veale, R. (2023). Pathways for naturalis-

tic looking behavior in primate I: Behavioral characteristics and brainstem circuits. *Neuroscience*, *532*, 133-163.

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), 4–4.

Veale, R., Murase, K., Watanabe, M., & Isa, T. (2024). Visual saliency predicts gaze during real-world driving task. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Veale, R., & Takahashi, M. (2024). Pathways for naturalistic looking behavior in primate ii. superior colliculus integrates parallel top-down and bottom-up inputs. *Neuroscience*.

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.

Whittaker, S. G., Cummings, R. W., & Swieson, L. R. (1991). Saccade control without a fovea. *Vision research*, *31*(12), 2209–2218.

Yoshida, M., Itti, L., Berg, D. J., Ikeda, T., Kato, R., Takaura, K., ... Isa, T. (2012). Residual attention guidance in blindsight monkeys watching complex natural scenes. *Current Biology*, *22*(15), 1429–1434.