Attentive Pointing in Natural Scenes Correlates with Other Measures of Attention

Daniel M. Jeck[1,2], Michael Qin[3], Howard Egeth[4], and Ernst Niebur[1,4,5]

[1]Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, MD

[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD

[3]Department of Biomedical Engineering, University of Connecticut at Storrs

[4]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD

[5]Solomon Snyder Department of Neuroscience, Johns Hopkins University, Baltimore, MD

April 13, 2017

Abstract

Finger pointing is a natural human behavior frequently used to draw attention to specific parts of sensory input. Since this pointing behavior is likely preceded and/or accompanied by the deployment of attention by the pointing person, we hypothesize that pointing can be used as a natural means of providing self-reports of attention and, in the case of visual input, visual salience. We here introduce a new method for assessing attentional choice by asking subjects to point to and tap the first place they look at on an image appearing on an electronic tablet screen. Our findings show that the tap data are well-correlated with other measures of attention, including eye fixations and selections of interesting image points, as well as with predictions of a saliency map model. We also develop an analysis method for comparing attentional maps (including fixations, reported points of interest, finger pointing, and computed salience) that takes into account the error in estimating those maps from a finite number of data points. This analysis strengthens our original findings by showing that the measured correlation between attentional maps drawn from identical underlying processes is systematically underestimated. The underestimation is strongest when the number of samples is small but it is always present. Our analysis method is not limited to data from attentional paradigms but, instead, it is broadly applicable to measures of similarity made between counts of multinomial data or probability distributions.

Keywords: pointing; tapping; salience; attention; probability density estimation; saliency map; natural scenes; fixations; interest points; probability distributions

1 Introduction

Factors influencing selective attention can notionally be separated into top-down and bottom-up influences. Top-down influences depend on the internal state of the observer, including his or her goals [*e.g.* Yarbus, 1967, DeAngelus and Pelz, 2009]. Bottom-up influences are factors that draw attention independently of any task and past experience with particular stimuli [*e.g.* Anderson et al., 2011]. For example, a bright flash in an otherwise still scene will usually attract attention [Yantis and Jonides, 1984]. The ability of parts of a visual scene to attract attention in a bottom-up fashion has been called the *salience* of this region [Koch and Ullman, 1985], a definition we adopt here.

While the definitions of top-down and bottom-up attention are clear, it is in practice difficult to dis-entangle their effects. For instance, observers who repeatedly perform tasks designed to measure bottom-up attentional effects may form expectations of what the next trial may be. These expectations will change their internal state and therefore add a top-down component to their responses. One of the goals of this study is to reduce such effects. Specifically, our goals are to:

- Introduce open ended self reports as a new experimental assay for selective attention and show that it can be measured efficiently using a pointing/tapping paradigm

- Develop a new experimental design in which each participant views only a small numbers of scenes. This reduces the contamination of bottom-up attentional effects by top-down expectations due to participants viewing similar stimuli many times

- Compare the results of this experiment with three other measures of attention and salience: fixations, interest points, and computed saliency

- Analyze the effects of sample size on estimating correlation between maps. The small number of samples from the pointing/tapping paradigm results in a statistical effect that causes the correlation between different maps to be systematically underestimated. We will clarify the influence of finite numbers of samples on the correlation between maps

1.1 Determining bottom-up saliency from human behavior

There are several methods that allow researchers to characterize items or regions that observers direct their attention to. One very influential approach has been visual search. Search for targets that differ from distractors by one of several low-level features (*e.g.* luminance, color, orientation contrast) takes a (generally short) time that is nearly independent of the number of distractors in the display [Egeth et al., 1972, Treisman and Gelade, 1980]. In contrast, targets that could be distinguished from distractors only by combinations of such features require search times that increased roughly linearly with the number of distractors [Treisman and Gelade, 1980, Egeth et al., 1984]. These and related results were fundamental in the construction of computational models for visual search [Wolfe et al., 1989, Wolfe, 1994, 2007] and for saliency determination and attentional selection [Niebur and Koch, 1996, Itti et al., 1998, Itti and Koch, 2001].

Given past success in utilizing features that promote efficient search, it is tempting to continue using visual search as a way to test models of visual salience. However, search tasks are limited in their applicability to measuring salience because participants are typically informed about the types of images they are about to see (*e.g.* "an image in which there is a single target and many distractors"), and the target and distractors are often described before the task begins. This information generates top-down influences that are likely to interact with bottom-up selection mechanisms. Even when participants are only told to look for a unique target, without being informed how it will differ from other objects ("odd-man out" tasks), they are still being informed about the structure of the image. It is then difficult to decide whether the participants find the target due to its bottom-up saliency features, or because of its uniqueness [Bacon and Egeth, 1994]. Results therefore may reflect a mixture of bottom-up (saliency) and top-down components of unknown composition.

This concern applies also to measurements of salience where participants give their subjective assessment of which of two stimuli is more salient [*e.g.* Nothdurft, 2000]. These experiments require that participants know that a stimulus will appear made up of oriented bars where two of them (one to the left and one to the right of fixation) will differ from the rest. As with search tasks, this information potentially biases the response of the participant. Indeed Nothdurft refers to needing additional concentration (clearly a top down process) to make difficult salience assessments. Furthermore, even if participants are not informed explicitly about the nature of the visual scene they are observing, the process of performing a task many times will likely give them information about what to expect.

While top-down influences can probably never be excluded entirely, our goal in this project is to reduce them. One possible way to mitigate top-down influences is to use "overt attention" in a free viewing task as an indicator for covert attention. In this approach, introduced by Parkhurst et al. [2002] and used in many subsequent studies [for a review see Borji and Itti, 2013], observers look at images (or videos) which can be natural or abstract scenes while their eye movements are tracked. Areas of the scene that are fixated are taken to be attended, a conclusion supported by findings from Deubel and Schneider [1996] that visual discrimination performance is enhanced at saccade targets. In the absence of a specific task ("free viewing"), it seems reasonable to assume that at least for the first few images, and for the first few fixations in these images, observers let themselves be guided by the visual input, rather than by some more complex strategy. This assumption becomes less plausible, however, the longer the sequence of images becomes and the longer the duration becomes that observers look at any given image. Indeed, Parkhurst et al. [2002] found that the agreement between eye fixation data and predictions of a purely bottom-up computational model of saliency decreased with viewing time/fixation number for a given image. It is not known whether the level of agreement depended on how many images had been viewed previously.

In principle it is possible to use the eye tracking method, with naïve participants viewing only a small number of scenes. In practice, the overhead of setting up an eye tracker system for each participant would make gathering fixation data for a small number of images per participant a very cumbersome task. We recruited 252 participants in this study, an order of magnitude more than participated in the latest saliency benchmark by Borji and Itti [2015], making eye-tracking each subject prohibitive.

To counteract this difficulty, we developed a novel experimental paradigm with the goal of gathering data from many participants where each participant only performed a small number of trials. The new paradigm is centered on showing subjects a short sequence of images and recording the response of each subject to each image. Some of the images are simple displays [similar to typical visual search arrays like those used by Treisman and Gelade, 1980] that are designed to test a specific hypothesis about what features of an image affect salience. Future work will discuss the structure of these images and the results gathered. Alternating with these images are natural scenes, the focus of this report. The goal in presenting these scenes to participants is to determine the extent to which salience as measured in our new experimental paradigm comports with salience data from previous studies. The natural scenes were therefore a subset of those used in a previous study [Masciocchi et al., 2009], and we will compare results obtained in our new paradigm with those from that study.

The data being compared here are attentional maps aggregated over a pool of participants. Such maps have been used in the study of salience extensively [Borji and Itti, 2013], and because they are population averages we can gather data to make attentional maps from a similar population without needing to gather new fixation data from the same subjects.

1.2 Reporting attended locations by pointing to them

Our new experimental paradigm for fast assessment of attentional selection was inspired by a study by Firestone and Scholl [2014] although those authors used a very different stimulus set and had a different motivation. The main idea is that, instead of recording eye movements, we ask participants to communicate their selections in a natural way by tapping on a screen with their (index) finger. Specifically, we ask the subjects to "tap the first place you look when the image appears." This instruction gives us a quick way to communicate in a non-technical manner that the participant should select the first attended location on the image, rather than an arbitrary point as requested by Firestone and Scholl [2014]. Even though instructions refer to where the participants look first, we do not attempt to determine whether any single individual is able to report their eye movements successfully. Instead, we are concerned with whether the population-level attentional maps we derive from the responses reflect previous measures of attention. We will validate our method by comparing these maps on when gathered for the same set of images.

We view this method of obtaining attentional maps as an alternative read-out of attention consisting of two (possibly interacting) components: self-report, and manual selection by finger tapping. Self reports have previously been taken as valid assessments of attentional selection when reporting attended locations in an experiment [*e.g.* Nothdurft, 2000]. Responding by tapping allows participants to indicate any location on the screen, rather than a pre-defined set of locations via a key press, or a less easily quantified verbal report. While it has been shown that planning manual movements can draw attention independently of eye movements [Jonikaitis and Deubel, 2011] in carefully controlled experiments,it is much more common for eye movements to guide hand movements when no experimental restrictions are in place [Fisk and Goodale, 1985, Neggers and Bekkering, 2000], minimizing the probability that a manual read out interferes with the self-report. Self reports also allow for the possibility of participants reporting the location of their covert attention rather than the location where they fixate, which may differ.

From a practical point of view, the method we use to record pointing behavior makes it a very fast, intuitive and simple process for collecting large amounts of selection data from a large and diverse participant population. Images were presented on an electronic tablet, and participants were instructed to tap on the first location that they looked at in the image, allowing for easy and precise recording of tap locations. In addition to allowing us to gather data from a large number of participants, the process reduces the information the participants were likely to have about the nature of the stimulus. We could then compare the responses of these relatively uninformed subjects to previously obtained measures of salience.

1.3 Limitations due to map estimation

We will follow the approach by Masciocchi et al. [2009] for computing correlations between different selection responses over the image. In that study, participants were asked to select interesting points on an image with a mouse. The distribution of selected points on the image was then interpreted as an estimate of the "interest map" internal to the participants that generated the data. Similarly, the distribution of recorded fixations from a free viewing task was turned into an estimate of a "fixation map." Both were compared with computed saliency maps. In the present study, we will introduce a third set of human response maps, defined by the pointing/tapping locations which we call "tap maps."

When comparing any two of these estimated maps, their measured correlation is determined by the nature of the two tasks and data types, as well as the amount of data collected to form the estimate. As we show in Section 2.3.3, the finite amount of collected data biases the computed correlation between maps toward zero. We develop a bootstrap procedure to estimate how large the bias would be if the two maps were drawn from the same underlying distribution. This procedure gives us insight into how correlated the data types could be and helps determine which comparisons between maps may benefit from further data collection.

2 Methods

All methods were approved by the Johns Hopkins Institutional Review Board and carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Alpha for all significance tests was set to 0.05. All data and code used for the analysis described in this section are available at https://github.com/dannyjeck/Attention-maps-comparison.

2.1 Apparatus, participants, and procedure

Participants were 252 passers-by on the Johns Hopkins University Homewood Campus (151 female; see Figure S1 for demographic information). They were approached by the experimenter and asked if they were interested in performing a short psychology experiment. If they answered in the affirmative, they were given instructions, as follows.

Participants were asked to give their gender (male/female) and age group (18-22, 23-30, 31-40, 41-50, and 51+). On a tablet computer (Apple Computers, iOS 8.3 operating system, screen 9.7" with 1024 × 768 resolution), participants were then shown a white screen with two small black squares (see Figure 1), which we call the initialization screen. They were informed that tapping on either one of the squares would bring up a test image, and were instructed, "When the image appears, tap the first place you look."

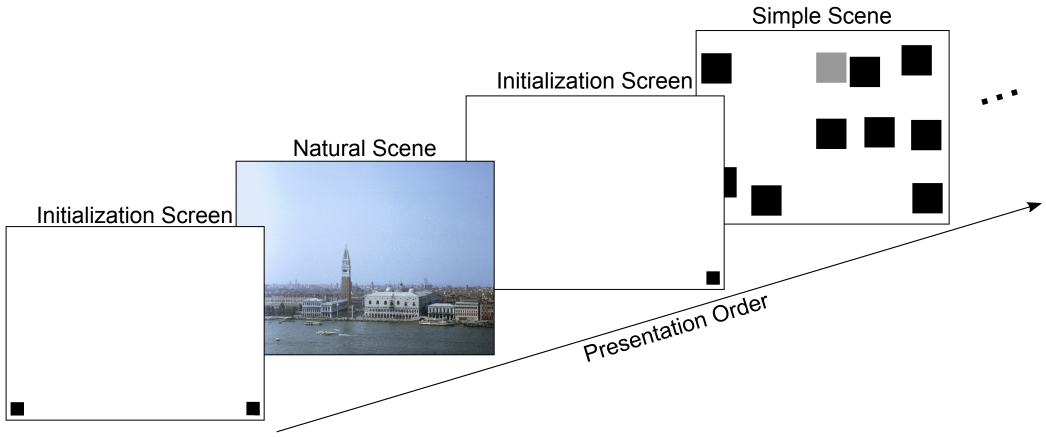

Figure 1: Experimental procedure. The rectangles represent an approximation of what was shown to participants on the tablet screen. First, they saw an initialization screen and tapped on either of the small black squares at the bottom. This brought up a test image which alternated between natural scenes and simple scenes. They then tapped on the test image at a place of their choosing which was, according to instructions, the first place they looked at when the test image had appeared. Tapping position and reaction time were collected, the initialization screen reappeared, and the cycle re-commenced.

After the participant had tapped first the initialization screen and then the location selected by him or her on the test image, the latter was immediately replaced by the initialization screen, and the cycle recommenced. This sequence of events continued until all images had been shown, with participants responding at their own pace. The position of the tap on the test image and the time between the taps on the initialization screen and on the test image were recorded. Test images strictly alternated between a natural scene and a simple scene consisting of colored squares on a white background, see section 2.2 and Figure 1. Each participant saw a total of 12 images of which the first always was a natural scene.

2.2 Stimuli

The stimulus set consisted of 48 natural scenes and 30 simple scenes. The natural scenes were taken from a previous study by Masciocchi et al. [2009] in which participants performed two tasks. One was free-viewing the scenes while their eye movements were recorded. In the other task, participants clicked with a mouse on locations on the scenes that they considered the most interesting; these locations were called "interest points." The size of the original images was 640 × 480, and they were resized for our purposes using MATLAB's (The MathWorks, Inc., Natick, MA) default image resizing function to fit the 1024 × 768 resolution of the tablet screen. Out of the four image classes in the Masciocchi et al. [2009] study we only used two, consisting of images of buildings and landscapes. Out of this set of 50 images, we randomly removed two to make the total number of natural scenes a multiple of six (the number of natural scenes each participant saw). The chosen 48 images were then separated into eight groups of six. The natural scenes for each participant rotated through these groups of six, such that every eighth participant saw the same six natural scenes. These scenes were presented in randomized order and always alternated with the simple scenes. No participant saw the same image twice.

The simple scenes consisted of a white background with randomly placed colored or gray-level squares, as shown in Figure 1. For the purposes of this study, they only served to interrupt the sequence of natural scenes and to decrease potential interactions between tapping locations on subsequent natural scenes. We note that the strict alternation of simple and natural scenes may allow participants to develop an expectation of the *type* of the subsequently presented image (simple or natural). Neither simple nor natural scenes are, however, predictive in any way about the *contents* of the next presented image, therefore no information about salient locations in an upcoming image is predicted by the sequence of images. Furthermore, no prediction is possible until at least one repetition has occurred, *i.e.* the second natural scene, which applies to one-third of the data collected.

2.3 Data Analysis

2.3.1 Correlations between maps

Selections of image areas by human observers (fixations, interest points, and taps) were first transformed into maps of the same dimension as the images. Computing the pairwise correlations between such maps as well as between the maps and the results of computational models of salience provide a measure of similarity between the different data collection methods and the models used. We reduced the resolution of the maps by binning the data. The reduction in resolution mitigates the possibility that fixations, taps, or interest selections that are near to each other are being counted as entirely distinct, though this is not the case for responses near the edge of the selected bins. We chose a 12×16 grid to tile the image (for an example see Figure 2B), therefore, each bin covers 64×64 image pixels. We chose this level of reduction in resolution since it is comparable to the eye tracker error used in obtaining fixation data [see Masciocchi et al., 2009, for details] and also roughly matches the size of a human finger pad when collecting tapping data. We also analyzed a coarser image resolution to examine the effects of resolution on the different correlations, results are shown in Figure 5. Similar findings between these two bin sizes confirm that the results are robust to bin size selection.

Tap maps were generated by weighing each tap on the appropriate image equally and binning them as described above. Interest maps were generated from from the data of Masciocchi et al. [2009], by taking each subjects first interest selection, the most interesting point per the instructions in the experiment, with each subject weighed equally. Fixations maps were generated by weighing each fixation by its duration. We also compared the distributions of fixations, interest points, and taps with saliency maps that were generated from the Itti et al. [1998] computational model of saliency at the same resolution.

Here we analyze the relationships between four processes: the three unknown processes, F generating fixation data, I generating interest point selections, T generating taps, and the known process S generating computed salience. If we assume each subject response is independent, then for a specific image, each unknown process can be described by a multinomial probability distribution (similar to a dice roll) from which data are drawn. We indicate the image number by adding a subscript to the process. For instance, for the k-th image I_k is a distribution from which each new interest point selection (by a different participant) is drawn. When we gather data, we are able to form estimates of these processes $\hat{F}_k$, $\hat{I}_k$, and $\hat{T}_k$ by computing the fraction of data points that fall in each bin for the k-th image. Since we are estimating a multinomial distribution using counts of the data, the resulting estimates of the rate of responses falling in a given bin are unbiased. However, as we will show in Section 2.3.3, the correlation values in comparing these maps are biased. Finally, as S is a known computational model, there is no need to form estimates of this process.

The measured covariation between any two processes P and Q on the k-th image, indexed in their horizontal and vertical dimensions by (i, j), with M bins total is,

$$\begin{aligned} C(\hat{P}_k, \hat{Q}_k) &= \frac{1}{M} \sum_{i,j} \hat{P}_k(i,j)\hat{Q}_k(i,j) - \frac{1}{M^2} \sum_{i,j} \hat{P}_k(i,j) \sum_{i,j} \hat{Q}_k(i,j) \\ &= \frac{1}{M} \sum_{i,j} \hat{P}_k(i,j)\hat{Q}_k(i,j) - \frac{1}{M^2} \end{aligned} \tag{1}$$

where the last equality holds because $\hat{P}_k$ and $\hat{Q}_k$ are probability distributions and therefore sum to unity.

The Pearson correlation coefficient R between estimates $\hat{P}_k$ and $\hat{Q}_k$ is then computed as,

$$R(\hat{P}_k, \hat{Q}_k) = \frac{C(\hat{P}_k, \hat{Q}_k)}{\sqrt{C(\hat{P}_k, \hat{P}_k)}\sqrt{C(\hat{Q}_k, \hat{Q}_k)}} \tag{2}$$

This quantity can vary between $R = -1$ for perfectly anticorrelated data and $R = 1$ for perfectly correlated data. We compare its value against two hypotheses, discussed in the following two subsections, 2.3.2 and 2.3.3. We refer to the average correlation coefficient over all images by dropping the subscripts in the argument.

2.3.2 Null hypothesis: Correlations reflect no differences between images

We consider first the (null) hypothesis that the contents of specific images do not affect the participants' responses. Under this hypothesis, for instance $R(\hat{F}_i, \hat{T}_i)$, the correlation between the fixation map from image i and the tap map from the same image is drawn from the same distribution as $R(\hat{F}_i, \hat{T}_j)$, the correlation between the fixation map from image i and the tap map from image j, for all i and j. We can approximate this null hypothesis distribution using a bootstrap technique to compute correlations between two types of maps (*e.g.* tap maps and fixation maps) using permutations of the image orders. Note that under this null hypothesis, image contents can still exert systematic influences on the selections but these influences do not differ systematically between different images. Therefore, the hypothesis includes correlations due to influences like center bias, "photographer's bias" (systematically placing objects of perceived importance in specific locations in the image), similarities due to similar image content, or other spatial preferences in common between participants. The null hypothesis does, however, exclude correlations caused by salient features of specific images.

2.3.3 Hypothesis: Correlations are limited by sampling error

At the other extreme, even for strong influences of image contents on correlations, estimating correlation from noisy estimates of the true processes generating the data create a bias in the measured correlation between any two types of maps. We illustrate this effect in a simple example. Consider two very simple one-dimensional identical distributions $P_k = Q_k = [0.5, 0, 0.5]$. If we draw an infinite number of samples from these (identical) distributions and use equation 2 to compute the correlation between the measured estimates, we obtain $R(\hat{P}_k, \hat{Q}_k) = 1$, as expected. But now consider the case of finite numbers of samples, and in the extreme, that only one sample from each distribution is drawn. Then, the estimate of the each distribution will either be $[1, 0, 0]$ or $[0, 0, 1]$. If they are the same, then $R(\hat{P}_k, \hat{Q}_k) = 1$ but if they are different $R(\hat{P}_k, \hat{Q}_k) = -\frac{1}{2}$. Therefore, the expected correlation is $\frac{1}{4}$. This bias towards zero will be non-zero for any finite number of samples drawn.

We want to gain an intuitive understanding of the bias in correlation for the unknown distributions underlying our data that is caused by the limited number of samples drawn. For this purpose, we developed a procedure in which we resample one of the maps with the same number of data points measured in the other to approximate how correlated the data could be under the hypothesis that the underlying processes were identical. Let P_k and Q_k be two processes with $\hat{P}_k$ estimated using n_P data points and $\hat{Q}_k$ estimated using n_Q data points, and let $n_P > n_Q$. First we select the type of map with the most data points, $\hat{P}_k$, and treat it as a perfect estimate of its underlying process. We then draw n_Q data points from $\hat{P}_k$ (with replacement) and compute a surrogate, $\tilde{P}_k^Q$. The tilde is used to indicate that the value is a resampling of the data from $\hat{P}_k$ and the superscript indicates the source of the number of data points used in the resampling. We then compute $R(\hat{P}_k, \tilde{P}_k^Q)$, the correlation between the surrogate data and the original map (see Figure 2C). For example, if the two maps in this procedure were fixations and taps and there were more fixations than taps, we would draw (with replacement) a number of surrogate data points from the fixation data set that was the same as that of recorded taps, and compute R between the surrogates and the original fixation map, $R(\hat{F}_k, \tilde{F}_k^T)$. For the reasons discussed in the previous paragraph, this value will be less than unity and it provides an intuitive estimate for how much the sampling error biases the measured correlations, $R(\hat{F}_k, \hat{T}_k)$. This procedure of generating surrogates and correlating with the original data can be repeated many times to refine the estimate of the bias in the correlation measurement under this hypothesis and to build a distribution against which to perform a hypothesis test (Figure 2D). We call this hypothesis the "sample error hypothesis," which assumes that a non-unity correlation measurement is due entirely to finite sample size. We note that this hypothesis is not truly an upper bound on the measured correlation (see Section 4.2 for a counterexample). We also note that, while this hypothesis is technically a null hypothesis against which we perform statistical tests, for the sake of clarity we will reserve the name "null hypothesis" for the hypothesis described in Section 2.3.2.

All resampling procedures were repeated with 1000 surrogates compared against the original.

2.3.4 Population averages

We analyzed the mean correlations between types of maps (*e.g.* taps and fixations) across all images (see Figure 3), which, as before, we denote by dropping the image number subscript. For example $R(\hat{F}, \hat{I})$ is the correlation between measured fixation and interest data averaged over all images. Similarly the average correlation under the assumption that the underlying distributions are actually identical (being the distribution of the interest data, which is the larger data set) and sampled with the number of fixations is given by $R(\hat{I}, \tilde{I}^F)$. The distributions of the null hypothesis differ between the combinations of maps but are identical for all image pairs of a given combination, *e.g.* Fixation and Interest maps in Figure 3B. Since many correlation values are averaged and we are measuring the difference between two mean values, hypothesis testing against the null becomes a two-sample Z-test. When testing against the sample error hypothesis we also perform a two-sample Z-test (see Supplementary Section S3 for validation of this method). Because both the final tests of significance average over all images and because the null and sample error hypotheses are relatively easy to reject (even though they are non-trivial), small p-values are expected. Beyond hypothesis testing, the mean correlation values provided by the null and sample error hypotheses also give points of reference against which we can compare the measured correlation values.

3 Results

We recorded 1510 taps from 252 participants (151 female; see Figure S1 for demographic information). The median of the reaction time (RT), defined as the time from tapping on the initialization screen to tapping on the test image, was about 1.4 seconds. Reaction times were skewed to the right (mean 1.6 seconds). We did not analyze RTs in detail because our data collection system did not allow precise control of the timing of image presentation. Data collection was completed after seven days of full time data collection.

3.1 Fixations vs. Interest Points

Aggregate results of our analysis for all images are shown in Figure 3. First, we re-analyzed the data from the Masciocchi et al. [2009] study with our methods. The analysis confirmed their result that interest and fixation data are correlated beyond the null hypothesis, $R(\hat{F}, \hat{I}) = 0.53$, Z-test $p = 1.3 \times 10^{-73}$; see Figure 3A. In addition, we now extend their results by showing that sufficient data was collected in that study so that the correlation under the sample error hypothesis between interest and fixations is very high,

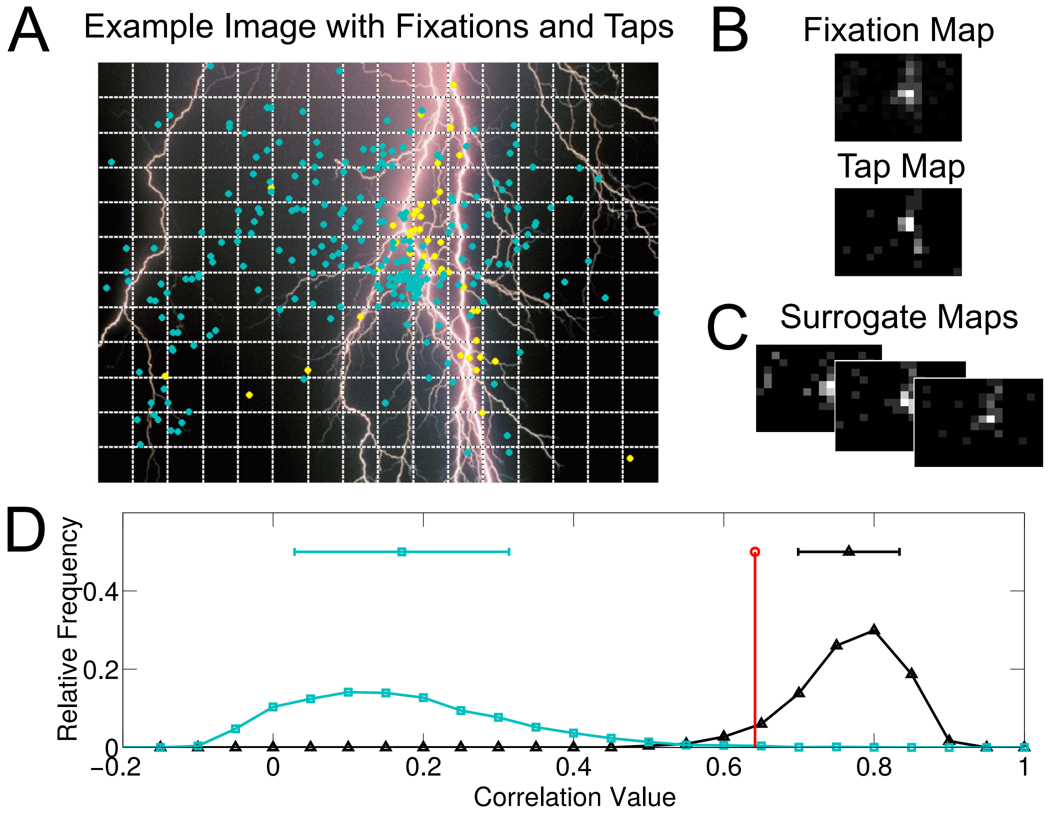

Figure 2: Data analysis method. (A) Example image overlaid with collected fixation points (blue dots) and tap points (yellow dots), and grid lines used to bin the data. (B) Corresponding fixation map and tap map. Both maps are binned in a 12×16 grid, with each bin showing the average of 64×64 pixels. (C) Surrogate maps generated from the fixation data used to approximate sampling error in the correlation between the fixation and tap data, see text. (D) Comparison of the measured value (red) to the histograms of the null hypothesis (blue) and the sampling error hypothesis (black). Means and standard deviations of the distributions generated from the null hypothesis and the sampling error hypothesis are shown above the distributions. For this image, fixation data and tap data correlate more than predicted by the null hypothesis ($p = 0.002$), and cannot be distinguished from predictions of the sampling error hypothesis ($p = 0.11$).

$R(\hat{F}, \tilde{F}^I) = 0.98$, indicating that the measure of correlation $R(\hat{F}, \hat{I}) = 0.53$ likely has very little bias. Differences between fixation and interest maps were not due to sampling error, Z-test $p = 1.5 \times 10^{-74}$.

3.2 Fixations vs. Computed Saliency

For the comparison of fixations and computed saliency from the Masciocchi et al. [2009] study (see Figure 3B) we found that the measured correlation exceeded the null hypothesis, $R(\hat{I}, S) = 0.19$, Z-test $p = 1.6 \times 10^{-16}$. Correlation under the sample error hypothesis is low for this comparison, $R(S, \tilde{S}^I) = 0.58$, though clearly higher than the measured correlation, Z-test $p = 6.8 \times 10^{-23}$.

3.3 Interest Points vs. Computed Saliency

We also compared interest points and computed saliency from Masciocchi et al. [2009], see Figure 3C. We found that the measured correlation exceeded the null hypothesis, $R(\hat{F}, S) = 0.30$, Z-test $p = 1.1 \times 10^{-18}$. Here the correlation under the sample error hypothesis is much lower than unity, $R(S, \tilde{S}^F) = 0.55$, indicating a potential bias in the measured correlation, though again higher than the measured values, Z-test $p = 9.5 \times 10^{-67}$.

3.4 Fixations *vs.* Tap Points

In the remaining three panels of Figure 3 we compare the correlations between the tap data collected in the present study with other attentional selection quantities. Correlations between fixation and tap data are shown in Figure 3D. The correlation level is similar to that between fixations and interest points in the Masciocchi et al. [2009] study, $R(\hat{F}, \hat{T}) = 0.45$, and it is again significantly above the null hypothesis ($p = 1.0 \times 10^{-39}$). Because fewer taps were collected than fixation points, the correlation under the sampling error hypothesis is $R(\hat{F}, \tilde{F}^T) = 0.64$. This is still significantly above the measured value ($p = 6.5 \times 10^{-16}$) but substantially below unity, indicating that the correlation may be substantially biased by the limited amount of data gathered.

It is unclear whether gathering more data would cause the measured correlation to increase or not. It may be that the "true" tap map T (which would be obtained if unlimited amounts of data were collected) is less diffuse than the measured fixation map $\hat{F}$, in which case the measured tap map $\hat{T}$ is a good estimate of the T map and the measured $R(\hat{F}, \hat{T})$ value is close to $R(F, T)$. Alternatively, the T map could be much more correlated with fixations than our measured map, in which case gathering more data will increase the correlation. We can say with high confidence that $R(F, T)$ is less than unity and greater than 0.41 (two standard errors below $R(\hat{F}, \hat{T}) = 0.45$).

We investigated the relationship between $R(F, T)$ and $R(F, I)$ further by computing $R(\hat{F}, \hat{T})$ and $R(\hat{F}, \hat{I})$ with subsets of the data collected for $\hat{T}$ and $\hat{I}$. We did this by drawing a number of data points without replacement from the tap data and interest data, and forming new estimates of the tap and interest maps. These were then correlated with $\hat{F}$ to qualitatively see whether the correlations $R(\hat{F}, \hat{T})$ and $R(\hat{F}, \hat{I})$ are converging as data is collected and to compare the two measures when equal numbers of data points are gathered. Results for various sizes of subsamples (up to the number of taps and interest points gathered per image) are shown in Figure 4. It is seen that for equal numbers of data points, $R(\hat{F}, \hat{T})$ and $R(\hat{F}, \hat{I})$ track each other closely, with both correlations increasing approximately logarithmically (about linearly in the semi-logarithmic plot) with the number of data points. For example, $R(\hat{F}, \hat{T}) = 0.44$ and $R(\hat{F}, \hat{I}) = 0.46$ when 29 interest points/taps are used per image. This is the largest number of taps available for all images. The number of data points available for fixations is larger than for taps and it can be seen that for much larger numbers (above ≈ 100), $R(\hat{F}, \hat{I})$ starts to plateau. The observation that $R(\hat{F}, \hat{I})$ plateaus agrees with our previous analysis that $R(\hat{F}, \hat{I})$ has very little bias since $R(\hat{F}, \tilde{F}^I)$ is nearly 1 and the asymptotic value in Figure 4 approaches the mean of $R(\hat{F}, \hat{I})$ shown in Figure 3B, about 0.53.

3.5 Interest Points vs. Tap Points

Tap data was also found to be significantly correlated with interest point data beyond the null hypothesis, $R(\hat{I}, \hat{T}) = 0.50$, $p = 1.2 \times 10^{-58}$, and correlation under the sample error hypothesis was significantly higher than the measured value, $R(\hat{I}, \tilde{I}^T) = 0.85$, $p = 1.3 \times 10^{-34}$, Figure 3E. The difference between $R(\hat{I}, \tilde{I}^T)$ and $R(\hat{F}, \tilde{F}^T)$ indicates that there is some difference between interest points and fixations that can not be explained by the smaller number of tap data. Despite drawing the same amount of data (the number of tap points) from the interest maps as we did from the fixation maps, the correlation under the sample error hypothesis is higher for interest maps because they are more focused than fixation maps (*i.e.* participants selected interest points in tighter clusters than was found in their fixations). Therefore, these clusters can be estimated more accurately with a smaller amount of tap data than for the more diffuse fixation maps.

3.6 Tap Points vs. Computed Saliency

Finally, saliency maps computed from the Itti et al. [1998] model were compared against the tap data and found to correlate beyond the null hypothesis, $R(S, \hat{T}) = 0.21$, $p = 4.3 \times 10^{-15}$, though not significantly below the sample error hypothesis, $R(S, \tilde{S}^T) = 0.25$, $p = 0.075$. This relatively low value of $R(S, \tilde{S}^T)$ is obtained because the computed saliency maps were relatively diffuse.

3.7 Coarse Scale Analysis

We also repeated the above analysis using fixation, interest, tap and salience maps at a coarser 3×4 resolution (the coarsest resolution possible with square bins). Results are shown in Figure 5. At this resolution all measured R values and resampled R values were higher, with measured R always falling between the null hypothesis and the sample error hypothesis (all $p < 0.05$). The level of measured correlation is thus dependent on the resolution used but the main results for the finer resolution hold. Because the measured correlations are still above the null hypothesis we can conclude that even for a very coarse grid, the image content is still informative beyond center bias, photographer's bias, or other structures common to a large fraction of images.

In summary, we found that tapping locations are correlated with the locations selected by each of the three measures considered previously: fixations, interest, and computed saliency [Masciocchi et al., 2009]. The null hypotheses of lack of correlation between tap locations and these three measures could all be rejected with high significance. Furthermore, we identified an important source of systematic downward shift (bias) of correlations between maps which is due to the finite numbers of selection points.

4 Discussion

4.1 A new experimental paradigm for quantitative characterization of attentional selection

We have developed a new experimental paradigm to evaluate what parts of an image attract the attention of observers. We do so by asking the study participants to report where they look and read out that report with a finger tap on the selected location. As far as we are aware, this is the first study in which open ended self-reports of attended locations are gathered. Unlike previous methods, this paradigm is particularly well suited to collecting data from participants who are not informed about the nature of what will be presented, mitigating top down effects related to expecting certain stimulus

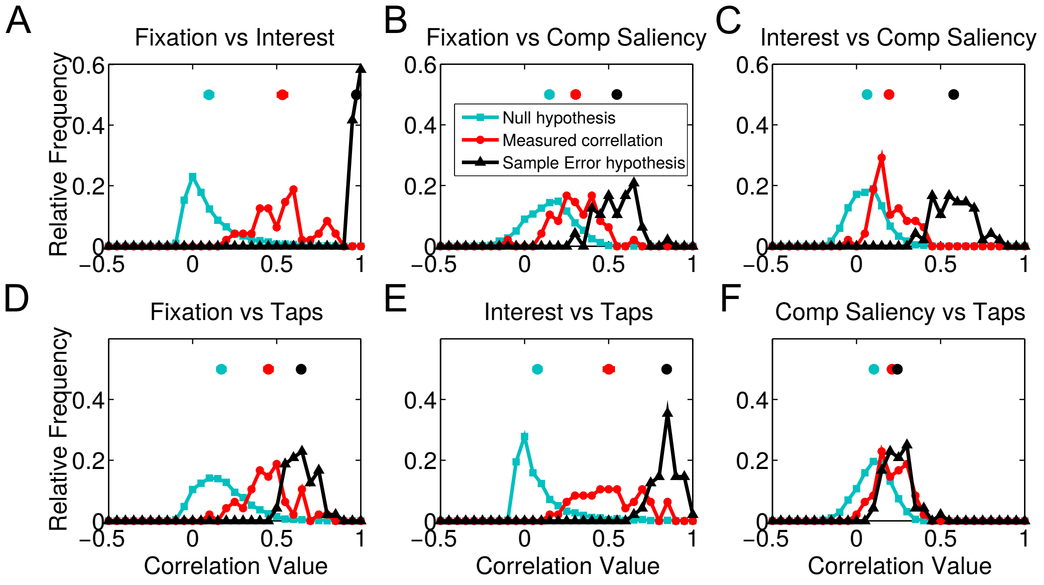

Figure 3: Aggregate results of natural scene analysis at 12 × 16 resolution. Each subplot shows a distribution of measured correlations between two types of maps compared against the null hypothesis and sample error hypothesis. Means of each distribution are shown above the histograms, with error bars indicating standard error given the 48 images used. Most error bars are smaller than the markers used. (A) Fixation and Interest maps. (B) Fixation and Computed saliency maps generated from Itti et al. [1998]. (C)Interest and saliency maps. (D) Fixation and Tap maps. (E) Interest and Tap maps. (F) Computed saliency and Tap maps. All measured averages are significantly above the null hypothesis ($p < 0.05$). All measured averages are below the sample error hypothesis ($p < 0.05$), with the exception of the comparison between computed saliency and tap maps ($p = 0.08$), panel F. The legend in panel B applies to all panels. For color figures see the online version of the article.

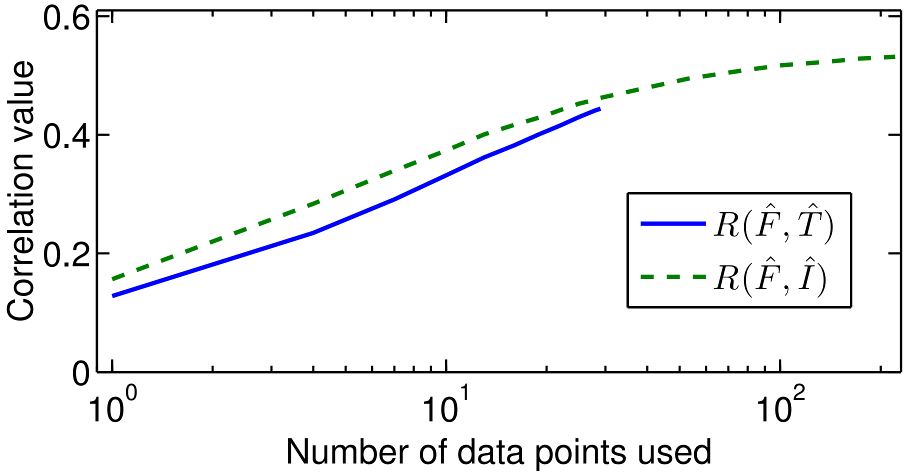

Figure 4: Comparison of $R(\hat{F},\hat{I})$ and $R(\hat{F},\hat{T})$ when using only a portion of the interest points and tap points. All fixation data was used to generate $\hat{F}$ for all simulations. 100 Simulations were performed for each number of data points. Standard error is less than line width. For color figures see the online version of the article.

types. We therefore interpret this new paradigm as a supplement to existing paradigms (free viewing, visual search, *etc.*) that can used to reduce top-down expectations that might bias participants' performance. Due to the simplicity of the experimental design, we were able to gather data from 252 subjects in seven days of data collection.

Pointing with a finger (similar to tapping a location) is a very natural and universal human behavior [Kita, 2003] which already appears during infancy, at about one year of age [Leavens et al., 2005, Tomasello et al., 2007]. The purpose of finger pointing is typically to direct attention (either that of the tapping person or more commonly that of another person) towards a specific part of the world. This behavior is thus often a direct, voluntary expression of attentional selection. It is more closely related to guiding the attentional direction of others than eye movements, although eye movements can also be used for directing attention in certain situations. While the term "overt

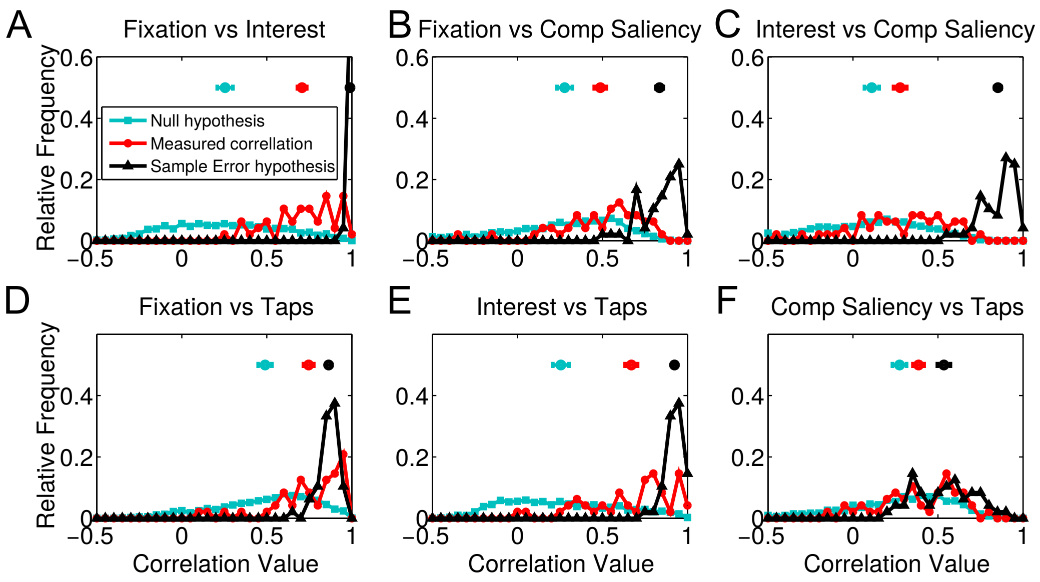

Figure 5: Aggregate results of correlation analysis at coarse resolution, when images were divided in a 3×4 grid. Symbols as in Figure 3. For color figures see the online version of the article.

attention" is traditionally used for eye movements (because they make the outcome of the covert attention process visible to the outside), pointing can therefore be seen as another form of overt attention, one that makes the outcome of the agent's attentional selection process explicit and instructs the observer to generate a "joint attentional frame" [Tomasello and Carpenter, 2007]. This strong connection with attentional selection makes this process not only attractive by itself, for the purpose of deducing the outcome of the covert selection process, but also for comparison with other correlates of attention, like eye movements and conscious selection of interesting parts of a scene. It thus complements the classical eye tracking method [Yarbus, 1967, Parkhurst et al., 2002] and the selection of interest points [Masciocchi et al., 2009].

The high levels of correlation between the four measures used in this study (fixations, interest points, taps and computed salience; see Figure 3) support the conclusion that the tapping paradigm is a valid measure of salience. For instance, the high correlation between taps and fixations ($R(\hat{F},\hat{T}) = 0.45$) indicates that the taps are capturing an aspect of salience seen in previous fixation studies. In fact, the value of $R(\hat{F},\hat{T})$ is likely biased downwards by the limited sample size,like all the correlations between maps. We have shown that if the fixations and taps were *perfectly* correlated, given the available number of data points the sampling error would still only result in a correlation coefficient of $R(\hat{F},\tilde{F}^T) = 0.64$. See Section 4.2 for further discussion of the sample error hypothesis

There are further factors that are expected to reduce the correlation of the measurements between taps and fixations, bolstering our result. The set of participants, screen, image resolution, and viewing conditions all varied between paradigms, and the outdoor conditions of the tap experiment allowed for multiple sources of possible distractions, including other passers-by. The fact that we find significant correlations in the presence of all of these confounding variables indicates that the responses given by participants are robust to a variety of low-level manipulations even though the measured correlations are likely decreased by these effects. Our finding suggests that attention is deployed based on invariant representations that are shared by the various participants and invariant to changes in viewing conditions.

Another difference between paradigms was the duration of presentation. While the tapping paradigm may be considered deliberative, the fixation data we used (from Parkhurst et al. [2002] and Masciocchi et al. [2009]) were gathered over a five second viewing period for each subject, more than three times the median reaction time during the tapping experiment (1.4 seconds). Free viewing periods of five second duration are in common use also for fixation datasets such as the widely used CAT2000 dataset [Borji and Itti, 2015]. Note that for the tapping study, the reaction time includes the time after the subject has decided where to tap, the movement of the hand, as well as the (relatively short) delay between the tap on the initialization screen and the presentation of the image. We therefore estimate that the majority of subjects performed three or fewer saccades before deciding where to tap. In principle, one could compare the tapping locations with only the first fixations from the studies that presented the same images Parkhurst et al. [2002], Masciocchi et al. [2009]. However, given the small number of participants in those studies, this analysis would not provide a meaningful map of fixated locations to compare against taps.

Finally, the process of making a hand movement may modify by itself the deployment of a participant's visual attention [Jonikaitis and Deubel, 2011, Baldauf and Deubel, 2008] thereby possibly changing the selected location. However, previous studies [Deubel and Schneider, 1996, Jonikaitis and Deubel, 2011, Baldauf and Deubel, 2008, Deubel and Schneider, 2003] all study conditions in which the reaching movements and saccades are planned in response to a cued location rather than indicating a salient stimulus. While more controlled research would be required to properly elucidate the interaction between manual selection and attention, we find it highly likely that the participant's selection is driven by their initial response to the image before the hand movement. If this were not the case, we would expect our measured correlations to be substantially lower.

In comparing the interest points and tap points (Figure 3 B-D), the results indicate that the correlation between our tap data and fixation data is approximately as strong as the correlation between fixations and interest points ($R(\hat{F}, \hat{T}) = 0.45$ *vs.* $R(\hat{F}, \hat{I}) = 0.53$). The correlation between interest and fixations is not subject to sample size bias to the same extent described above because the correlation under the sampling error hypothesis ($R(\hat{F}, \tilde{F}^I) = 0.98$) is so close to unity. Given these results, we speculate that the responses for the tap experiment lie somewhere in between the more involuntary fixation responses and the more deliberative responses given in the interest points task.

The level of correlation between taps and computed salience ($R(S, \hat{T}) =$

0.21) in the natural scenes was lower than previous findings indicated for other correlates of attentional selection. Masciocchi et al. [2009] found the correlation coefficients between fixations and computed salience to be $R = 0.32$, and between interest and computed salience to be $R = 0.37$ using slightly different methods. The results of Masciocchi et al. [2009] are in closer agreement with our low-resolution analysis, which found $R(S, \hat{T}) = 0.38$ and $R(S, \tilde{S}^T) = 0.53$. These results indicate that the salience model from Itti et al. [1998] which was used in both the previous study and this one captures a substantial aspect of the bottom up processes that influence attention. However given the low correlation value, it is likely that other aspects of those processes are not being captured.

Overall, our results show highly significant correlations between attentional selections executed by the oculomotor system [Parkhurst et al, 2002, and many other more recent studies; for a review see Borji and Itti, 2013] and by the skeletomuscular system. For the latter, this is the case both when conscious deliberation is encouraged [Masciocchi et al., 2009] and when it is discouraged (this study). Remarkably, these measures also correlate well with predictions of a very simple computational model of bottom-up attention [Itti et al., 1998]. Without doubt, this simple model has limitations, *e.g.* in the representation of objects [Einhäuser et al., 2008, but see Borji et al, 2013], even though they can be overcome at least partially by more sophisticated proto-object based models [Mihalas et al., 2011, Russell et al., 2014]. However, the fact that even a very basic model captures human behavior over such a large range of tasks illustrates the fundamental role of attentional selection for behavior.

4.2 Effects of sampling error on correlations

Another contribution of this study is a new way of analyzing correlations between maps of different types, such as fixations or taps, although our method should apply to many other kinds of maps. These maps are generated by accumulating many individual measurements into a “heat map,” which can be interpreted as an estimate of the probability distribution of the data. The measured correlation between the maps (*e.g.* $R(\hat{F}, \hat{T})$) and the estimates of those probability distributions (here $\hat{F}$ and $\hat{T}$)) will depend on both the underlying distributions (F and T) and the quality of the estimates. The differences between the true distributions are of scientific interest. For the case of the maps considered in this study, these differences may be useful in determining what aspects of a scene draw attention, and their correlation is useful in determining the validity of the tap experiment as a measure of salience.

Estimates of the true distributions based on finite amounts of data will, however, bias our estimate of the correlation. With an infinite number of data points, the true distributions could be measured to perfect accuracy. Given a fixed limited sample size, increasing the resolution of the maps increases the number of parameters in the distribution to be estimated and therefore decreases the accuracy. Similarly, if the true distribution is spread widely across the image, the accuracy of the estimate will be reduced much in the same way that, everything else being equal, the standard error of the mean for a distribution with high variance is greater than the standard error of the mean for one with low variance.

This source of bias in correlation measurements differs from the reduction (“attenuation”) in correlation described by Spearman [1904] when measuring the correlation between two signals in noise. While both effects bias the observed correlation towards zero, the underlying mechanisms are quite different between our effect and Spearman’s, making his method for correcting the bias inappropriate in our case. Spearman observed that the correlation between two processes is attenuated if noise is added to one or both of them, and in his 1904 study he developed a method to correct for the bias found in correlating noisy measurements. In contrast, in the effect described in the present study, no noise is added. The bias in the correlation here is due to the finite number of observations of the underlying distributions (for tap, fixation, and interest selection). In the example in Section 2.3.3 of the two simple distributions, the correlation is biased because we only sample from a small number of points (in the extreme case discussed, just one), but there is no noise in the samples. The two effects are independent, one could have one or the other or both, and each contributes its own bias to the total decrease of the correlation. For instance, while the bias due to the limited sample size described in Section 2.3.3 disappears if the sample size goes to infinity, this is not the case for the noise-induced attenuation effect discovered by Spearman [1904].

One may still be tempted to apply the method from Spearman [1904] to correct for the bias found in correlating noisy measurements of probability distributions. After all, the estimates of probabilities can be thought of as a measurement of the true distributions plus noise. However, the noise characteristics are entirely different in the present case. Spearman [1904] assumes independent identically distributed additive noise, while the estimation error resulting from drawing a finite number of samples from a multinomial distribution is dependent on the value measured and exhibits covariation between bins (since the error must sum to zero) Spearman's method is therefore not a valid solution to this problem.

Given the potential sources of error in estimating correlation, we have developed a simulation-based method (Section 2.3 and Figure 2) to compute the correlation between maps assuming that the true maps are perfectly correlated. Note that, although one might think that the correlation of a map with itself is an upper bound on the correlation of the map with other maps, even for finite numbers of samples, this is not the case. For a counter example, if $\hat{P} = [0, 1, 0]$ is measured with one sample, and $\hat{Q} = [0.3, 0.4, 0.3]$ is measured with (infinitely) many samples, then $R(\hat{P}, \hat{Q}) = 1$, but the expected value of $R(\hat{Q}, \tilde{Q}^P)$ is 0.1 because there is a probability of 0.6 that the single sample drawn from Q will be from either the first or last bin. In this case, the correlation is $-\frac{1}{2}$ because the peak in one distribution aligns with one of the two equal troughs in the second.

The use of Pearson correlation (R) is useful in gaining a qualitative measure of the similarity between the distributions. Overlapping peaks and troughs in distributions will result in positive R values. However, R is invariant to linear scaling. If one distribution is relatively uniform while another has high peaks and troughs, the R function may find them to be highly correlated so long as their peaks and troughs align. As such, the correlations measured in this study show that interest points, taps, and fixations all seem to fall on similar locations, though the distributions may have substantial differences under another metric.

The method of estimating the sampling error effect that we introduce is applicable to any correlation computation between estimates of a true distribution. In fact, the method can be extended to any metric of similarity between distributions or maps. For example, if Kullback-Leibler divergence (KLD) is believed to be a more appropriate metric of similarity, the sample error hypothesis can be used to generate surrogate data under the hypothesis that the two types of data are drawn from the same distribution. Then the KLD between the surrogate data and the original map can be used to determine the size of the sampling error effects.

We also note that there may be methods to reduce the bias in the measured correlation using a Jackknife procedure [Efron, 1982], though it is unknown to what extent such a procedure would introduce unwanted variance into the estimation procedure.

Acknowledgments

We acknowledge support through a Visual Science Training Program Fellowship to DJ. This research was supported by the Office of Naval Research under Grant N000141010278, and the National Institutes of Health under

Grant R01DA040990-01. We thank Matthew Babina for his comments on the manuscript. We gratefully acknowledge discussions with Prof. Daniel Naiman about statistical methodology.

Supplementary information

S1 Demographics

Detailed demographics are shown in Figure S1. Participants were passers-by on the Johns Hopkins University campus. No deliberate selection criterion was applied, except for (possibly unconscious) perceptions of approachability and whether the individuals seemed in too much haste to be likely willing to participate in the experiment. *Post-hoc* we noticed that gender groups were generally balanced, with the exception of the 23-30 age group in which female participants dominated for unknown reasons.

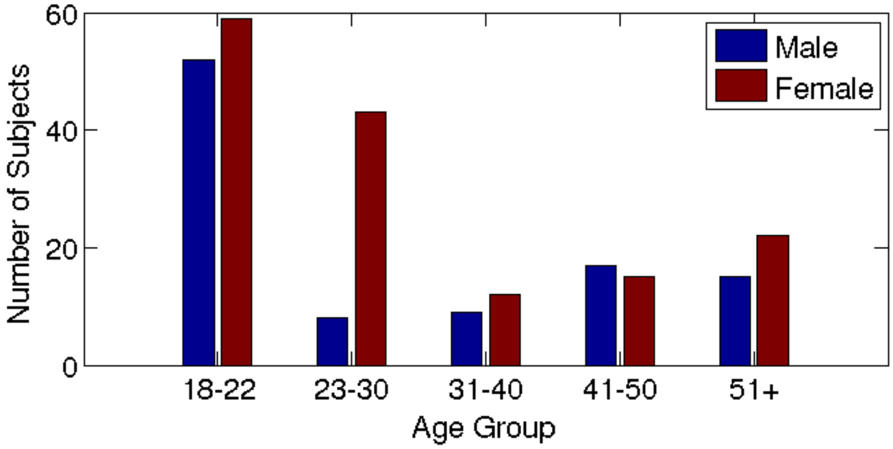

Figure S1: Demographics of the 252 participants.

S2 Error modes

The experiment had a number of error modes, as follows.

1. Two taps by one participant were lost when transmitting data from the tablet to the server.

2. Some participants would tap the black square on the right side of the initialization screen but the tablet registered the tap as being on the status bar (not visible to the participant) and did not process it within the experiment. This caused some confusion for some early participants before an image was presented, but no data was lost. Later participants were told to only use the square on the left, away from the status bar.

3. Some participants would accidentally tap either the test image or the initialization screen twice in rapid succession. 15 taps were recorded to take place within 400 milliseconds of another tap.

4. There was some variability between the loading times of images. Some seemed to consistently load more slowly than others. As mentioned in the main text, we did not analyze reaction times in detail for this reason.

5. One participant seemed to understand the instructions when starting the experiment, but this became doubtful while she performed the experiment. She tapped in a tight group on the right side of the screen. The possible reason was that she was English-challenged, something that was not apparent while she was recruited and instructed.

6. Some participants seemed to consistently take a very long time to complete the task.

Since all these error modes resulted in a very small number of possibly problematic taps, no exclusion criteria were defined before analyzing the data, none was excluded. All participants and taps are included in the analysis of the paper barring the two taps that were not recorded.

S3 Statistical validation

To validate our statistical approach we will first repeat our tests using a standard bootstrap technique, and then introduce the motivation and validation of the technique used in the main text.

A canonical bootstrap technique [Efron, 1982] draws samples with replacement from some empirical distribution to generate new samples. This is the way we generate the surrogate maps under the sample error hypothesis. A standard way to gather p-values is to generate surrogate samples under a null hypothesis and compare a measured value to those samples. Consider as an example the sample error hypothesis that $R(\hat{F}, \hat{T})$ is a sample from $R(\hat{F}, \tilde{F}^T)$. Let N be the number of samples from $R(\hat{F}, \tilde{F}^T)$ that are drawn, and n be the number of those samples that satisfy

$$R(\hat{F}, \hat{T}) \geq R(\hat{F}, \tilde{F}^T)$$

We can then generate a valid p-value as

$$p = \frac{n+1}{N+1} \tag{3}$$

Here, the $+1$ in the numerator and denominator arise because when hypothesis testing we assume the null is true, and therefore the measured value of $R(\hat{F}, \hat{T})$ is also part of the null hypothesis.

We computed p-values using equation 3 on the data shown in Figure 3D using 1000 samples drawn from the sample error hypothesis. The measured value of $R(\hat{F}, \hat{T})$ did not exceed any of the 1000 surrogate correlation values. We repeated this analysis for each of the sample error hypotheses shown in Figure 3 and obtained the same result. All p-values are therefore equal to 1/1001. This includes the case of $R(S, \hat{T})$, which had a p-value above 0.05 in the main text.

A hypothesis test is considered valid if, when the null hypothesis is true, the rate of getting a p-value below a threshold α is less than or equal to α [Casella and Berger, 2002]. This is true if the distribution of p-values under the null hypothesis is uniform, or if the left side of the distribution is lower than a uniform distribution (in which case it is also called a conservative test). To further validate the simple bootstrap test from equation 3, we generated 1000 p-values when $R(\hat{F}, \hat{T})$ is replaced with a sample from $R(\hat{F}, \tilde{F}^T)$ (*i.e.* assuming that the sample error hypothesis is true) to show that the distribution of p-values is uniform. This is, indeed, the case, as shown in Figure S2A.

While these results confirm the validity of our hypothesis test with the chosen $\alpha = 0.05$, we were curious how confident we can be that our results hold for stricter choices of α. We could choose to generate more samples from the sample error hypothesis, however these are computationally expensive and unreasonably large numbers of samples would be needed to obtain the low p-values we measure. An alternative approach is to use a closed-form approximation of the distribution of interest and then compute the p-values using that approximate distribution. Because the correlations we test are all averages over many images, we chose a Gaussian approximation. The associated hypothesis test is therefore a two-sample Z-test. In order to validate the approximation we must ensure that p-values generated under the null hypothesis are valid. To do so we repeat the processing used to generate Figure S2A, but now we compute the p-values using the Z-test. The positive slope of the resulting distribution (shown in Figure S2B) indicates that the test is valid, and indeed conservative, with (much) fewer than 50 of the 1000 p-values below the threshold of 0.05 that would be expected under a uniform distribution.

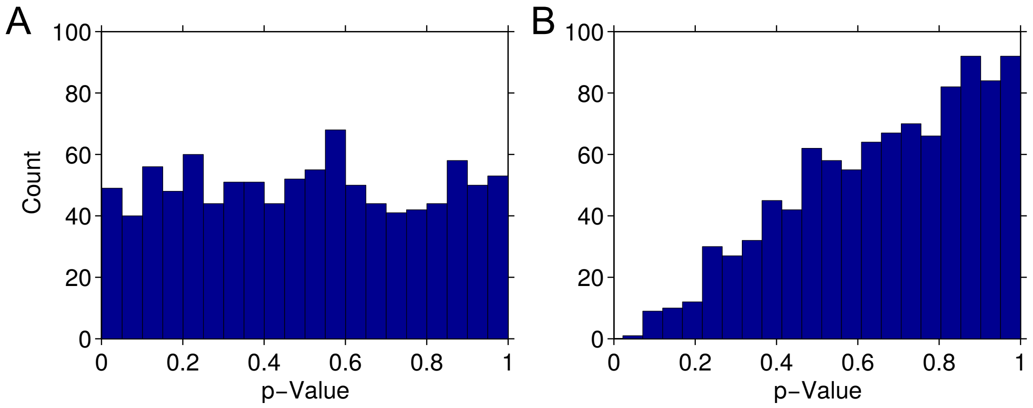

Figure S2: Histograms of 1000 p-values under the null hypothesis (A) under the empirical p-value from equation 3, and (B) under the Gaussian assumption from the main text.

References

B.A. Anderson, P.A. Laurent, and S Yantis. Value-driven attentional capture. *Proc. Nat. Acad. Sci., USA*, 2011.

W. F. Bacon and H. E. Egeth. Overriding stimulus-driven attentional capture. *Perception & Psychophysics*, 55:485–496, 1994.

Daniel Baldauf and Heiner Deubel. Visual attention during the preparation of bimanual movements. *Vision Research*, 2008.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

Ali Borji, Dicky N Sihite, and Laurent Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of vision*, 13(10):18, 2013.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Marianne DeAngelus and Jeff B. Pelz. Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6-7):790–811, August 2009. ISSN 1350-6285. doi: 10.1080/13506280902793843. URL http://www.tandfonline.com/doi/abs/10.1080/13506280902793843.

Heiner Deubel and Werner X Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1837, 1996.

Heiner Deubel and Werner X. Schneider. Delayed saccades, but not delayed manual aiming movements, require visual attention shifts. *Annals of the New York Academy of Sciences*, 2003.

Bradley Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

H.E. Egeth, R.A. Virzi, and H. Garbart. Searching for conjunctively defined targets. *J. Experimental Psychology*, 10(1):32–39, 1984.

Howard Egeth, John Jonides, and Sally Wall. Parallel processing of multielement displays. *Cognitive Psychology*, 3(4):674–698, 1972.

W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vision*, 8(14):1–26, 2008.

Chaz Firestone and Brian J Scholl. "Please tap the shape, anywhere you like": Shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological science*, 25(2):377–86, February 2014. ISSN 1467-9280. doi: 10.1177/0956797613507584. URL http://www.ncbi.nlm.nih.gov/pubmed/24406395.

J. D. Fisk and M. A. Goodale. The organization of eye and limb movements during unrestricted reaching to targets in contralateral and ipsilateral visual space. *Experimental Brain Research*, 1985.

L. Itti and C. Koch. Computational modelling of visual attention. *Nature Neuroscience*, 2:194–203, 2001.

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

Donatas Jonikaitis and Heiner Deubel. Independent allocation of attention to eye and hand targets in coordinated eye-hand movements. *Psychological science : a journal of the American Psychological Society / APS*, 2011.

Sotaro Kita. *Pointing: Where language, culture, and cognition meet.* Psychology Press, 2003.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol.*, 4:219–227, 1985.

David A Leavens, William D Hopkins, and Kim A Bard. Understanding the point of chimpanzee pointing epigenesis and ecological validity. *Current Directions in Psychological Science*, 14(4):185–189, 2005.

C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur. Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, 9(11):1–22, October 2009.

S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur. Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proceedings of the National Academy of Sciences*, 108 (18):7583–8, 2011. PMC3088583.

S F W Neggers and H Bekkering. Ocular gaze is anchored to the target of an ongoing pointing movement. *Journal of Neurophysiology*, 2000.

E. Niebur and C. Koch. Control of selective visual attention: Modeling the "where" pathway. In D. S Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–808. MIT Press, Cambridge, MA, 1996.

Hans-Christoph Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40(10-12):1183–1201, 2000. ISSN 00426989. doi: 10.1016/S0042-6989(00)00031-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0042698900000316.

D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of visual selective attention. *Vision Research*, 42(1):107–123, 2002.

A. F. Russell, S Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94: 1–15, 2014.

C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101, 1904. ISSN 0002-9556. doi: 10.2307/1412159. URL http://www.jstor.org/stable/1412159.

Michael Tomasello and Malinda Carpenter. Shared intentionality. *Developmental science*, 10(1):121–125, 2007.

Michael Tomasello, Malinda Carpenter, and Ulf Liszkowski. A new look at infant pointing. *Child development*, 78(3):705–722, 2007.

A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. PMID: 7351125.

J. M. Wolfe. Guided search 2.0 – a revised model of visual search. *Psychonomics Bulletin & Review*, 1(2):202–238, 1994.

J.M. Wolfe. Guided search 4.0. *Integrated models of cognitive systems*, pages 99–119, 2007.

J.M. Wolfe, K.R. Cave, and S.L. Franzel. Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychology*, 15: 419–433, 1989.

S. Yantis and J. Jonides. Abrupt visual onsets and selective attention: evidence from visual search. *J Exp Psychol Hum Percept Perform*, 10:601–621, Oct 1984.

A.L. Yarbus. *Eye Movements and Vision.* Plenum Press, New York, 1967.