PLOS ONE

RESEARCH ARTICLE

Rapid visual categorization is not guided by early salience-based selection

John K. Tsotsos*, Iuliia Kotseruba, Calden Wloka

Department of Electrical Engineering and Computer Science, York University, Toronto, ON, Canada

* tsotsos@eecs.yorku.ca

OPEN ACCESS

Citation: Tsotsos JK, Kotseruba I, Wloka C (2019) Rapid visual categorization is not guided by early salience-based selection. PLoS ONE 14(10): e0224306. https://doi.org/10.1371/journal.pone.0224306

Editor: Mudassar Raza, COMSATS University Islamabad, Wah Campus, PAKISTAN

Received: February 20, 2019

Accepted: October 11, 2019

Published: October 24, 2019

Data Availability Statement: Two image datasets were used in this paper. One image dataset is from Thorpe et al, originally purchased from the Corel Stock Photo Library (https://www.amazon.com/Corel-Stock-Photo-Library-2/dp/B000V933GI). The ground truth masks for this data created by the authors have been made publicly available at the URL below. The other image dataset is from Potter et al. The authors confirm they have permission to share the Potter et al image data and ground truth masks created by the authors. These data are available from the Center for Open Science (https://osf.io/cemkw/). Additionally, all algorithms used in

Abstract

The current dominant visual processing paradigm in both human and machine research is the feedforward, layered hierarchy of neural-like processing elements. Within this paradigm, visual saliency is seen by many to have a specific role, namely that of early selection. Early selection is thought to enable very fast visual performance by limiting processing to only the most salient candidate portions of an image. This strategy has led to a plethora of saliency algorithms that have indeed improved processing time efficiency in machine algorithms, which in turn have strengthened the suggestion that human vision also employs a similar early selection strategy. However, at least one set of critical tests of this idea has never been performed with respect to the role of early selection in human vision. How would the best of the current saliency models perform on the stimuli used by experimentalists who first provided evidence for this visual processing paradigm? Would the algorithms really provide correct candidate sub-images to enable fast categorization on those same images? Do humans really need this early selection for their impressive performance? Here, we report on a new series of tests of these questions whose results suggest that it is quite unlikely that such an early selection process has any role in human rapid visual categorization.

Introduction

The current dominant visual processing paradigm in both human and machine research is the learned, feedforward, layered hierarchical network of neural-like processing elements, with a history stretching from Rosenblatt's Perceptrons [1], Fukushima's Cognitron [2] (and subsequent Neocognitron [3]), to Rumelhart & McClelland's Parallel Distributed Processes [4], LeCun & Bengio's Convolutional Neural Networks [5], and to Krizhevsky, Sutskever, and Hinton's Deep Neural Networks [6]. The goal of each of these models and systems was to explain or emulate the effortless ability of humans to immediately perceive content in images. Tsotsos [7] termed this *immediate vision* and laid out the computational difficulty of the task as well as key elements of how brains and machines might defeat its combinatorial nature.

Our everyday experience tells us that vision feels immediate: we simply open our eyes and the world is there, fully formed before us and ready for our interactions. There is no perceptible time delay or inner 'turning of wheels'. It is well-documented over several decades that humans can recognize visual targets with remarkably short exposure times, with the seminal

the study have publicly available code with URLs provided in the Supporting Information.

Funding: This research was supported by grants to the senior author (JKT) from the following sources: Air Force Office of Scientific Research USA (FA9550-18-1-0054) (https://www.wpafb.af.mil/afrl/afosr/), Office of Naval Research USA (N00178-16-P-0087) (https://www.onr.navy.mil/), The Canada Research Chairs Program (950-231659) (http://www.chairs-chaires.gc.ca/home-accueil-eng.aspx) and Natural Sciences and Engineering Research Council of Canada (RGPIN-2016-05352) (https://www.canada.ca/en/science-engineering-research.html). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

works including Potter & Levy [8], Potter & Faulconer [9], Potter [10], Thorpe et al. [11], and more recently Potter et al. [12]. The short exposure times (the shortest being 13ms) and subsequent fast responses (150ms of neural processing required for yes-no answers to categorize an image) led theoreticians to conclude that there was no time available for any processing other than a single pass through the visual hierarchy in the feedforward direction [13].

To be sure, there are a variety of models and theories that add feedback and recurrence to such hierarchical networks from Fukushima [14] to Hochreiter & Schmidhuber [15] and Sutskever [16] and more, but with respect to the main thrust of this paper, these do not detract from the main conclusions here because they address tasks different from rapid visual categorization and thus, generally, would be inconsistent with the observed time course of human categorization performance.

Still, the computational requirements to fully process a whole scene seem daunting [17] and many suggested that there must be some sort of optimizing action to reduce computational load occurring along that feedforward path. Within this paradigm, the processing of visual saliency has been suggested to have this specific role, namely reducing computational load via early selection [18]. Early selection was thought to reduce the information that must be processed to enable very fast visual performance by determining a spatial region-of-interest (ROI) on which further analysis should be performed. In the Koch & Ullman formulation [18], a saliency map is computed early in the visual processing stream and represents point-wise stimulus conspicuity (contrast between a point and its local surround). A winner-take-all competition selects the most conspicuous location (point) and the features at the selected location are routed to a central representation for further processing. It is important to note that Koch and Ullman viewed saliency as a method for predicting the most useful image locations for processing in recognition or similar higher level tasks; eye movements were not considered as an outcome of saliency computation in their paper. Inhibition of that selected location forces a shift to the next most conspicuous location when the algorithm is run again. Koch & Ullman's early selection idea seems to have been motivated by Feature Integration Theory [19] in that it provided a mechanistic version of its operation, specifically, the selection of a focus within the master map of locations. It shares much with the Broadbent's Early Selection model, which was based on human auditory behavior [20]. In the first stage 'physical' properties (e.g. pitch) would be extracted for all incoming (auditory) stimuli, in a 'parallel' manner and in the second stage, psychological properties, beyond simple physical characteristics (e.g. meaning of spoken words) would be extracted. This second stage was more limited in capacity, so that it could not deal with all the incoming information at once when there were multiple stimuli (having to process them 'serially', rather than in parallel). A selective filter protected the second stage from overload, passing to it only those stimuli which had a particular physical property, from among those already extracted for all stimuli within the first stage. Many criticized Broadbent's early selection idea and proposed alternates including late selection schemes [21], [22], [23], [24], and attenuation schemes [25].

Early implementations of saliency computation did indeed produce points of maximum conspicuity that were found useful for machine vision [26], [27], [28], [29]. As algorithms evolved, they moved more towards salient region or object proposals (for reviews see [30], [31], [32]). Fixation-based algorithms are typically validated by how well they match human fixation points (even though eye movements were not part of the original experimental work nor are they the only manifestation of attentional behaviour), while salient object detection algorithms are validated by how well the regions they produce overlap with ground truth object outlines or bounding boxes. The number of different saliency conceptualizations and implementations now is in the hundreds [33]. Models based on deep learning methods have

also recently embraced this early selection idea in the hopes that their already impressive success can be improved further [34], [35].

Many high-profile models of human visual information processing, appearing over the past 3 decades, include some variant of early selection within a feedforward visual processing stream [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. These not only claim biological inspiration but also biological realism. That is, the authors claim that their processing strategies actually reflect the brain's visual processing strategy. Often, it is difficult to evaluate such claims. For example, Yan et al. [47] claim that their evidence regarding V1 representation during an orientation singleton task, where a monkey learns to make an eye movement to a singleton oriented bar, reflects bottom-up salience and go as far as to assert that V1 computes a bottom-up saliency map. Their experiments do not address whether any of the other representations throughout the visual cortex have a similar characteristic. Whether or not such representation exists in the brain has been addressed by many, all of whom find supporting evidence for a saliency map, including: superior colliculus [48], [49] [50]; LGN [51], [52]; pulvinar [53], [54], [55]; FEF [56]; parietal areas [57]. In each of these, the connection to a saliency representation is made because maxima of responses that are found within a neural population correspond with the behaviorally attended location. The Yan et al. work also draws their conclusion based on this observation. Could it be that they all do simultaneously?

It almost seems a straightforward inference that if V1 shows this characteristic, each visual area receiving feedforward input from V1 necessarily also shows it, and this continues through the visual hierarchy. The use of machine learning methods to demonstrate 'read out' of neural response only shows that it is possible to extract the necessary information from a neural population and not that this is the actual sole source of that information. Since the information that would lead to behavior is necessarily in the stimulus itself, any representation of that stimulus that is created in a sufficiently non-destructive manner necessarily also includes that same information. In making claims about a single locus for a saliency map, these authors fail to provide a mechanistic explanation for how behavior is generated directly from that representation and without any relevant influence from other brain processing areas. To demonstrate the existence of a saliency map in any representation, one must present evidence that some retinotopic visual area alone has a causal connection to behavior. It is also important to recall that eye movements, the behavior linked to saliency in Yan et al. [47], played no role in the original conceptualization of the saliency map.

Hypothesis

The present work examines one of the basic underlying features of all these models: that salience-based early selection within a feedforward visual hierarchy provides a spatial ROI on which further analysis is performed (i.e., the basic Koch & Ullman idea described above). Our motivation was born from the observation that at least one set of critical tests of this idea has never been performed. No one has tested saliency models on the stimuli that were used in the seminal experiments that supported the feedforward view. Do these algorithms really provide an accurate reduction of the visual search space to enable fast categorization? Here, the first question we ask is: If the computation of saliency occurs early in the feedforward pass through visual areas, and determines a location for further processing, does the first ROI determined by a saliency algorithm effectively point to the target?

If the question were to be answered in the affirmative, then when the images used in the seminal experiments are run through a saliency algorithm, the algorithm should yield a prediction for the target location that matches well with the ground truth. It would be reasonable to assume that a good prediction is followed by a correct categorization because a good

prediction identifies the target sufficiently well. That only the first selection of an algorithm is of interest is key: the temporal constraints provided by the experimental observations and which theorists used (e.g., Feldman & Ballard [13]) do not permit more than one selection.

After conducting this experiment, the results pointed us to a second question: is human rapid visual categorization guided by early selection? This led to a second experiment, this time examining human behaviour, where we manipulated the stimulus image set to discover the target location and extent needed for good categorization. While addressing these two questions, we also examined other issues, specifically, center bias in datasets and algorithm biological plausibility. Together, the natural conclusion is that early selection played no role in the key seminal experiments.

Methodology

There are several main elements that comprise our approach to the questions raised about early selection: the image data sets, the set of algorithms tested, the performance metrics used, the analysis of algorithm biological plausibility, and human experiments (approved by the York University Office of Research Ethics, certificate 2016-014 "Selective tuning approach to visual system attention executive") to examine performance to parafoveal stimuli from the image data sets. Each will be described in turn.

Before these descriptions, the overall logic of this argument and the role of these components is presented. If early selection via image saliency plays a role in rapid human visual categorization, then we hoped to find existing saliency algorithms, that generally have very strong performance on available benchmarks, that approached human performance in their ability to predict targets. If we succeeded, then the good algorithms would point to potential directions for how they might be improved and utilized for machine vision. We also wanted to answer the early selection question for human vision so needed to examine the algorithms not only for their accuracy but also whether they embodied basic biological constraints known to underlie human categorization behavior. We could thus say whether the human design inspiration was useful. When we realized that we were finding mostly negative results on these counts, we wondered whether humans needed any such early guidance at all and tested this by creating a full set of parafoveal stimuli for the same categorization task. The results on all of these experiments point to the facts that the development of salience algorithms is not yet at a point where human level performance can be expected and that humans might have no need of such early guidance for the original categorization tasks of the seminal experiments in any case. In order to answer all of the above questions we developed a set of metrics with which we compared algorithm and human performance.

Image sets and summary of original results

The following seminal experiments and image datasets were considered. Potter & Levy [8] examined memory for visual events occurring at and near the rate of eye fixations. Their subjects were shown sequences of 16 pictures, from 272 magazine photos, presented with rates of 0.5, 1, 2, 3, 4, 6, or 8 images per second. They concluded that rapidly presented pictures are processed separately for precisely the time each is in view and are not held with other items in a short-term memory. This was among the earliest works to demonstrate rapid categorization but the image set was unavailable for our use.

Potter and Faulconer [9] used 96 stimuli, half of these being line drawings and half words that represent objects in those line drawings. There were 18 categories of objects and within each, between 2 and 9 instances (e.g., food: carrot, pie; clothing: hat, coat; tools: pliers, hammer). Each stimulus was preceded and followed by a mask of random lines and pieces of

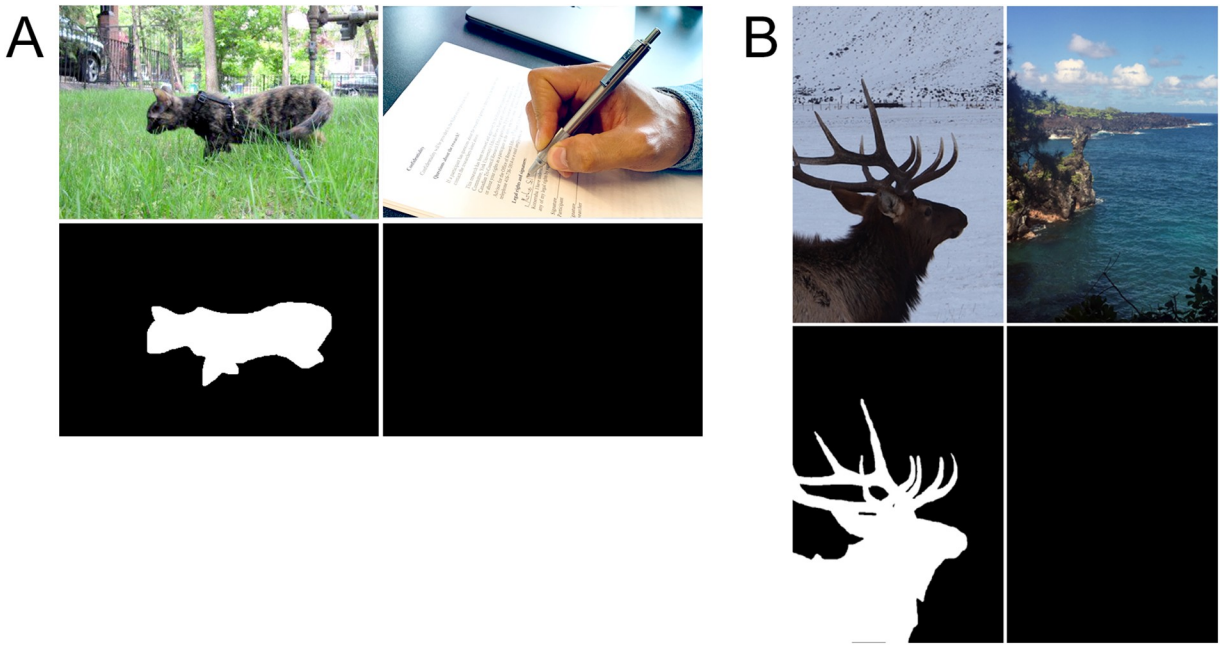

Fig 1. Sample categorization images. A: Sample images representative of those from the Potter dataset with target present (top left), no target present (top right), and corresponding ground truth masks (bottom row). In the experiment, participants were asked if they saw a particular target (here, "a walking cat"). Only those trials where the question was asked after the stimulus presentation were used. The bottom panels show the hand-drawn ground truth masks; no-target images have no ground-truth. B: Sample images, representative of those in the Thorpe dataset with target present (top left), no target present (top right), and corresponding ground truth masks (bottom row). In the experiment participants were asked if there is an animal in the image. The bottom panels show the hand-drawn ground truth masks; no-target images have no ground-truth.

https://doi.org/10.1371/journal.pone.0224306.g001

letters. Target information was provided to subjects in some trials before and, in others, after the stimulus. Stimuli were shown for 40, 50, 60 or 70ms. They observed that subjects needed 44ms exposure duration for the drawings and 46ms for the words to achieve 50% accuracy. We obtained these images and tested some of the saliency algorithms; however, we did not pursue these. The saliency algorithms all produced simply blurred versions of the line drawings and thus were not useful, likely due to their development being primarily based on natural images. This description is included here because it sets an early data point for fast categorization.

Potter [10] used an RSVP (Rapid Serial Visual Presentation) task with 16 photos of natural scenes. Subjects were either shown the target that they might find within the sequence in advance or were told its name in advance. Accuracy was over 70% after only 125ms of exposure. In a second experiment, she tested subject's memory. Subjects looked at a 16-image sequence of pictures without prior instruction and were then asked a yes-no question about what they had seen. Subjects required about 300ms exposure to achieve 50% accuracy. Regrettably, this dataset was also unavailable.

Fortunately, there was an alternate stimulus set that was available. Potter et al. [12] used an RSVP task of a series of six or 12 color pictures presented at 13, 27, 53 and 80ms per picture, with no inter-stimulus interval. Images were 300x200 in size, and there were 1711 images in total, 366 with target present. An example is shown in Fig 1A, with the corresponding hand-drawn ground truth mask (GTM). Participants were to detect the presence or absence of a target specified by a name (e.g., smiling couple) that was given just before or immediately after the sequence (in other words, subjects only had usable target expectations in half the trials). If

subjects reported a positive detection, they were then asked a 2-alternative forced choice (2AFC) question to see if they could recognize the target given a distractor. Detection improved with increasing duration (from 13ms up to 80ms) and was generally better when the name was presented before the sequence, but performance was significantly above chance at all durations, whether the target was named before or after the sequence. At the shortest exposure, prior knowledge seemed to provide no benefit at all. For the set of trials without prior expectations, the ones relevant to our study, performance of the 2AFC task ranged from about 67% to about 73% correct on the target-present trials. Performance when prior expectation was provided ranged from 75-85% accuracy. The results are consistent with feedforward models, in which an initial wave of neural activity through the ventral stream is sufficient to allow identification of a complex visual stimulus in a single forward pass. Potter and her colleagues generously provided this dataset for our work and confirmed that these stimuli were of the same type as used in [8] and [10].

Thorpe et al. [11] ran 'yes-no' categorization tasks. Subjects viewed color images (with none repeated during trials), and had to determine if an animal was present or not. The original images were 512x768 in size but downsized in Thorpe's experiment to 256x384 (which we used), totaled 2000 images, with 996 having target present. A representative example is shown in Fig 1B. There was no prior knowledge of types of animals and stimuli were taken from commercial images. They measured behavior plus ERP (Event Related Potential). Even though the duration of the image exposure was 20ms, subjects exhibited 94% average correctness. Prefrontal ERP activity diverged at 150ms after stimuli onset for 'yes' and 'no' responses, which means enough processing had been done in 150ms to decide if an animal is present or not. They concluded that sufficient processing must be occurring in a primarily feedforward manner. Thorpe and colleagues graciously provided the full original dataset for our research.

The two chosen experiments for our comparison are not identical and some justification as to why they are suitable for our test of feedforward saliency is in order. In the Thorpe et al. case, the processing path is direct and very much what we need to compare against; if feedforward saliency computation is part of human categorization performance it would definitely be part of the 150ms time period Thorpe et al. reported. The most direct other experiment for us to include would have been the Potter 1975 paper [10]. The closest we have to this is the Potter et al. stimulus set [12], confirmed to involve stimuli of the same type as the earlier paper. The detection component, which would reflect the same direct path as Thorpe et al., is present but was followed with an additional task. This means we should not compare time courses—and we do not. The detection task is reported using d' values in the main paper but also using percent correct in their supplementary material, which is what we used. Only the results for trials where there was no prior knowledge are relevant for our work. They show that for target-present trials, proportion correct improved as stimulus duration increased from 13 ms to 80ms, from about 60% to 73% (we use 73% as the performance mark), while for target-absent trials the mean correct was 75%. These values were for their 6 picture test; for their 12-picture test, the accuracy was similar. We stress that our tests do not impact the validity of any of the original experiments cited.

The tested algorithm set

To conduct our test, we chose 7 bottom-up fixation based saliency models. Each is referred to by the acronym in bold given here. Two are algorithms that represent a cross-section of classical methods: the most commonly used and cited model by Itti et al. (ITTI) [29] and the AIM model [58], a consistently high performing model in benchmark fixation tests. We also selected several recent algorithms which achieved high scores in the MIT benchmark [59] and

had publicly available source code (see S1 Text for details), namely Saliency in Context (oSALICON—the open source version) [60], Boolean Map based Saliency (BMS) [61], Ensembles of Deep Networks (eDN) by [62], RARE2012 [63], and DeepGaze II [64]. eDN, oSALICON, and DeepGaze II represent the class of saliency algorithms based on deep learning. eDN model is a set of shallow neuromophic networks selected via hyperparameter optimization for best performance on the MIT1003 saliency dataset [65]. The other two models rely on transfer learning from deep networks initially trained on object classification tasks (VGG-19 [66] in DeepGaze II and VGG-16 [66] in oSALICON) to achieve state-of-the-art performance on the MIT saliency benchmark. Finally, we also added the 'objectness' algorithm (OBJ) because the human experiments all involve categorization of objects [67]. All algorithms were used with default parameters and published implementations. Many algorithms use an explicit center bias typically expressed as a centered Gaussian distribution in order to improve performance (typically a gain of 2-3%). For those models the bias was disabled to enable a fair comarison. See S1 Text for implementation details. A 9th method was added for control purposes, which we refer to as CENTER. This places the point of interest at the center of the image regardless of image contents. We use $\mathcal{P}$ to denote the point of interest for all algorithms.

How one measures performance is very important especially when it involves direct comparison of human and machine output. In the absence of eye movements, the degree of overlap between a target and the region of high acuity in the retina is likely strongly correlated with human performance. As a result, some of our performance measures include this overlap. This seems easily justifiable by considering the details of human photoreceptor layout on the retina (see S2 Text for details). Observers in the original experiments were instructed to fixate the image center and there was no time for any change of gaze. Thus, if a sufficient spatial extent of the target lies within the observer's parafovea, it would seem that detection should be more likely correct.

If this assertion is appropriate, then image sets whose targets are strongly center-biased would lead to better categorization performance than those image sets with lesser bias. We thus created scatterplots for the target centroids of the two stimulus sets. This revealed a substantial center bias for the Potter set (Fig 2A) and a strong center bias for the Thorpe set (Fig 2B). Sure enough, Thorpe et al. reported higher human performance than Potter et al.

Performance metrics

Our purpose was to determine guidelines for how we measure algorithm correctness during our tests. We considered several different ways of developing these metrics and a brief description of their derivation follows based on the human performance levels presented in the previous section. Thorpe et al. observed a 94% accuracy in their experiment; 94% of the targets had a least 27% of their extent within the parafovea. Similarly, Potter et al. observed a peak accuracy of 73% and 73% of targets had at least 41% of their area within the parafovea. To be sure, there is no correspondence between the set of observed correct responses, either per subject nor collectively, and the set of ground truth masks identified with this analysis. But such a correspondence cannot be computed with the available data and our purpose was not to correctly determine this correspondence. Samples of this calculation are shown in Fig 3 (the example with the image of an elk (top row) demonstrates where this assumption may be inappropriate).

Using these estimates, which are admittedly coarse at best, we limited the image region where a saliency algorithm prediction would be considered as a valid ROI cue for human categorization. Note that this is not a measure of algorithm correctness in the manner usually used in benchmark tests [68]. The goal is to quantify how well saliency algorithms provide guidance

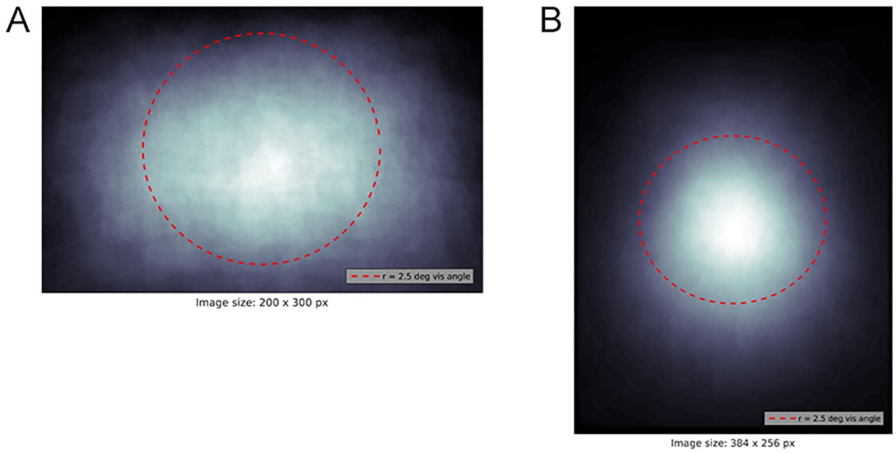

Fig 2. Target distribution in datasets. Two plots showing overlaid ground truth mask centroids for targets in the Potter (A) and Thorpe (B) datasets. Brighter pixels correspond to greater overlap between the masks. Parafovea (r = 2.5˚) is shown by a dashed red line. Mean of the distribution lies approximately in the center of the image. The area within 2.5˚ from the center of the image is calculated based on the following assumptions: the viewer is 57 cm away from the monitor, the monitor has 23" diagonal and resolution of 1920x1080.

https://doi.org/10.1371/journal.pone.0224306.g002

for the human visual system towards the task of accurate image categorization. In any case, these are only two of the performance measures; the other two have no similar approximate nature.

We thus decided on four separate ways of quantifying algorithm performance. A saliency algorithm's predicted first point of interest, $\mathcal{P}$, would be marked as correct if:

1. $\mathcal{P}$ is anywhere within the GTM;
2. $\mathcal{P}$ is within the GTM AND within the parafovea, anywhere (even if by one pixel);
3. $\mathcal{P}$ falls within both the GTM AND the parafovea AND at least 27% of the GTM (by area) lies within the parafovea. This reflects the reality of Thorpe et al. data and will be applied only for those stimuli;
4. $\mathcal{P}$ falls within the GTM AND within the parafovea AND at least 41% of the GTM (by area) lies within the parafovea. This reflects the reality of Potter et al. data and will be applied only for those stimuli.

When compared to the observed target layout characteristics, these are conditions which are very generous in favor of the algorithms. One additional point must be addressed. None of the tested algorithms include the capacity to accept prior instruction. The Potter et al. results we use as comparison are those without subjects receiving prior guidance while those of Thorpe included uniform guidance for expected category. Although this might appear to lead to an unfair comparison, we note that the performance in the Potter et al. experiment without prior expectation was roughly 10% lower than with prior guidance [12]. As a result, we can reasonably assume a similar decrease in the Thorpe experiment, and as will be seen, this will not affect the overall conclusion.

To better understand the appropriateness of these choices, we conducted a sensitivity analysis for each of the algorithms and this is shown in Fig 4. The plots show how the percentage of points of interest (P) within the ground truth masks and parafovea gradually decreases depending on how much of the GTM (by area) is inside the parafovea. On each plot the point (0,0) corresponds to measure B (i.e. P is within the GTM and parafovea regardless of the

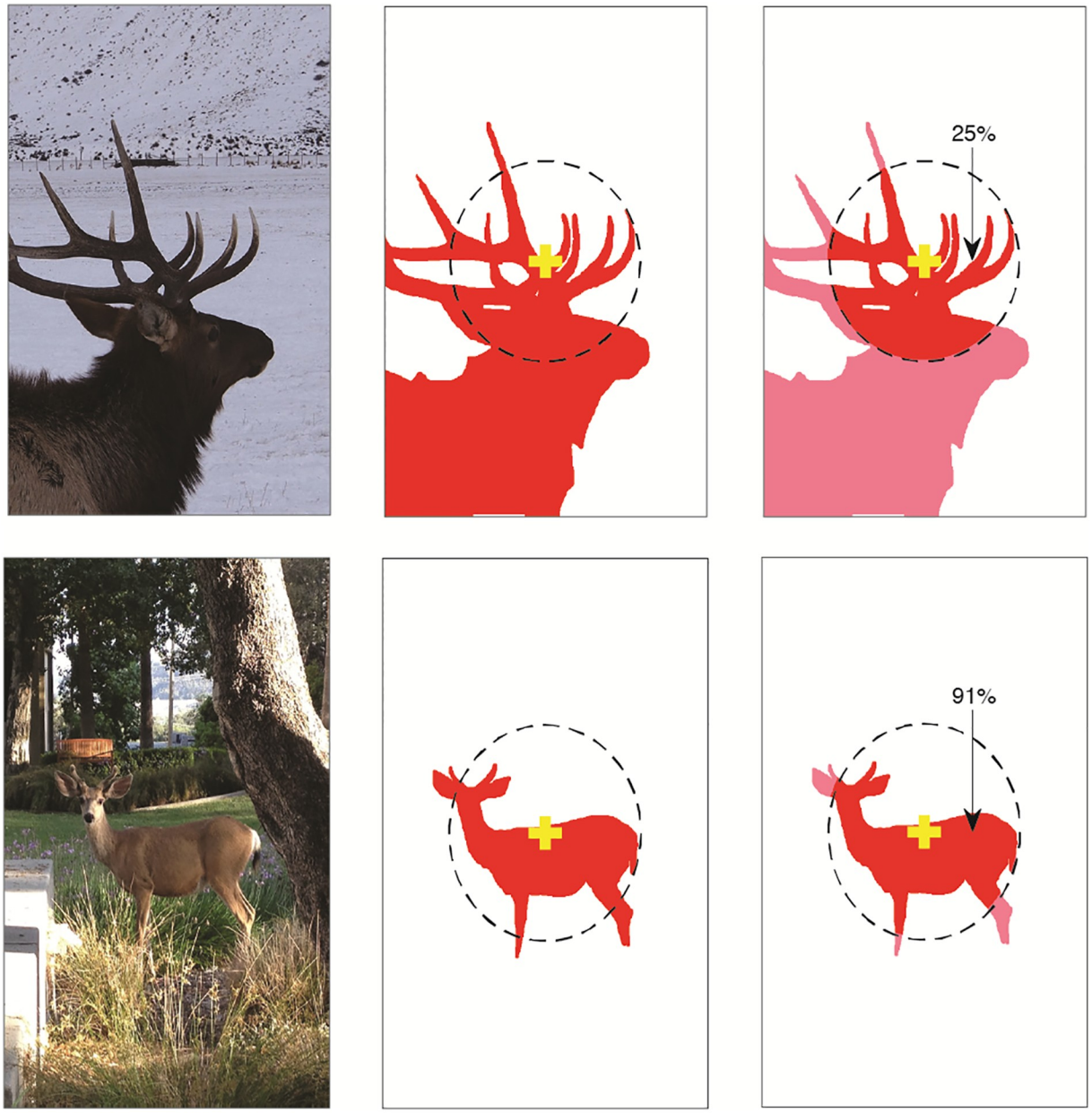

Fig 3. Visualization of performance measures. Two examples are shown. The original stimulus image is on the left with the subject's fixation point marked. The middle image gives the ground truth mask where if the saliency algorithm produces a predicted $\mathcal{P}$, our measure A will count a positive hit. The right image shows the extent of the human parafovea, the dashed circle, centered at subject's fixation and superimposed on the GTM. Measure B will count a prediction as a hit if it falls within the marked area. Measure C and D will count a prediction as hit if measure B is a hit and the percentage of GTM area within the parafovea is sufficiently large. For these examples, 25% of the elk GTM area lies within the parafovea while the deer is 91% within. The elk image would lead to a 'hit' only for measures A and B whereas all the measures would count the deer as a hit.

https://doi.org/10.1371/journal.pone.0224306.g003

amount of GTM within the parafovea). Dashed vertical lines show what amount of overlap corresponds to human performance in Potter and Thorpe experiments. The overlap of 27% corresponds to measure C and 41% to measure D shown in Fig 5 (C and D). This analysis makes clear that the threshold choices are indeed sensible and fair, and that they are not

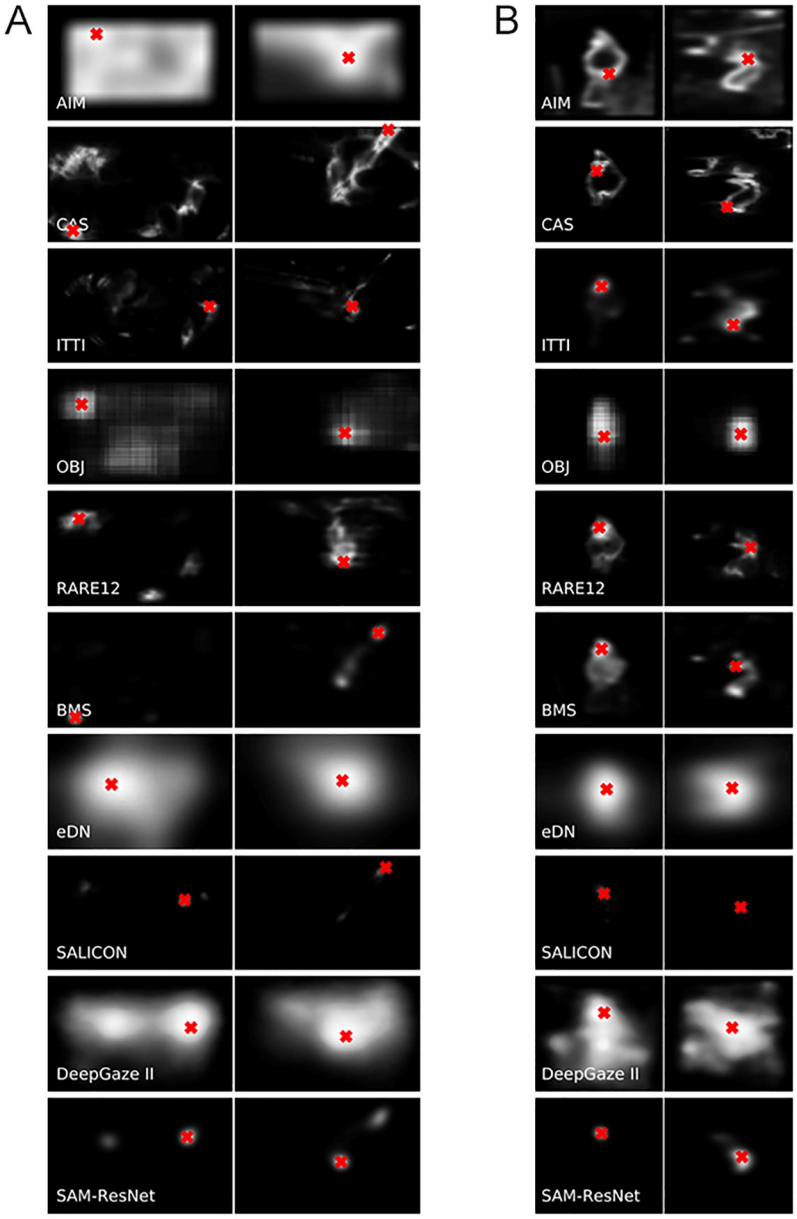

Fig 4. Sensitivity analysis of performance thresholds. The plots show how percentage of points of interest P within the ground truth masks and parafovea gradually decreases depending on how much of the GTM (by area) is inside the parafovea. On each plot the point (0,0) corresponds to measure B (i.e. P is within the GTM and parafovea regardless of the amount of GTM within the parafovea). Dashed vertical lines show what amount of overlap corresponds to human performance in Potter and Thorpe experiments.

https://doi.org/10.1371/journal.pone.0224306.g004

sensitive to small changes. Finally, as will be shown in the Results section 'c', these threshold choices are justifiable through human experimentation.

Algorithm biological plausibility

Although many computer vision algorithms take significant inspiration from the human visual system, few embody that inspiration in a biologically realistic manner. It is difficult to evaluate such algorithms as to their biological plausibility but there is at least one tool that can be used.

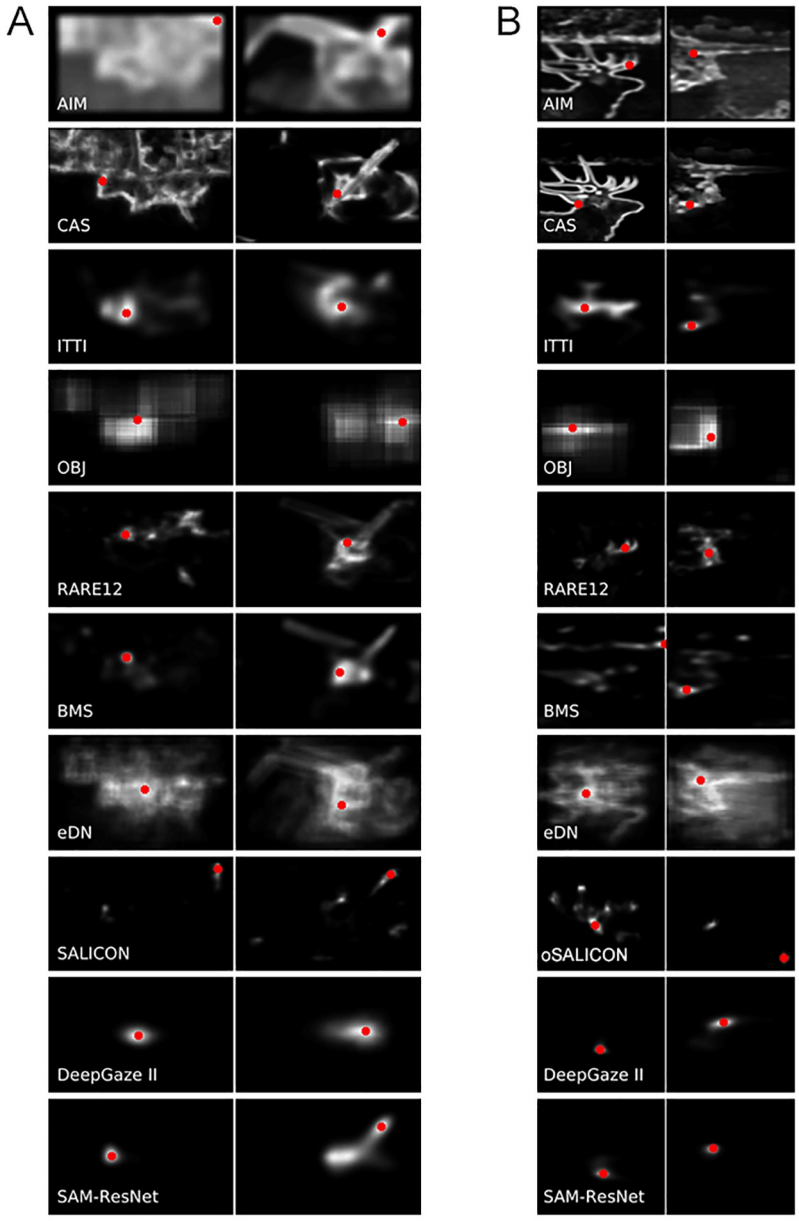

Fig 5. Examples of saliency maps. Saliency maps generated by the 8 saliency algorithms for sample images from the Potter (A) and Thorpe (B) datasets shown in Fig 1 with target present (left column) and no target present (right column). The red dot in each saliency map marks the global maximum found in the saliency map (the most likely first attended location predicted by the algorithm).

https://doi.org/10.1371/journal.pone.0224306.g005

Feldman & Ballard [13] explicitly linked computational complexity to neural processes saying "Contemporary computer science has sharpened our notions of what is 'computable' to include bounds on time, storage, and other resources. It does not seem unreasonable to require that computational models in cognitive science be at least plausible in their postulated resource requirements." They go on to examine the resources of time and numbers of processors, and more, leading to a key conclusion that complex behaviors can be carried out in fewer than one hundred (neural processing) time steps. These time steps were considered to be roughly the time it might take a single neuron to perform its basic computation (coarsely stated as a

weighted sum of its inputs followed by a non-linear transformation) and then transmit its results to the next level of computation, perhaps about 10ms. Thorpe & Imbert [69] also place similar constraints on processing time and numbers of layers suggesting that at least 10 layers of about 10msec per layer are needed. Combining this with Thorpe et al.'s observation that 150ms suffices for yes-no category decision, this constrains biologically plausible algorithms to those requiring no more than 15 or so layers of such computations. Since the algorithms we are testing do not deal with the full problem of categorization but only reflect the saliency computation stage, one might expect a much smaller time constraint, i.e., significantly fewer than 15 layers of computation.

Human categorization performance to parafoveal stimuli

If early selection, salience-based or otherwise, were important for human rapid categorization performance, then testing humans with images where only the region within an observer's parafovea (within 2.5˚ radius of point of fixation) would be revealing. Good performance would show early guidance is unnecessary. In other words, the current pre-determined fixation would suffice for good performance implying a shift in fixation would not lead to meaningful improvement. We took the original image set of Thorpe et al. and cropped each image to its parafovea content and tested subjects for categorization performance.

Results

Algorithm performance

Fig 5 shows one example saliency heat map from each of the two datasets for each algorithm, one for a target-present and one for target-absent (the images used are those shown in Fig 1).

The overall categorization performance data is seen in Fig 6, where parts A, B, C and D correspond to the four performance measures described earlier, respectively.

This test reveals several results:

1. Using the most generous measure, A, several algorithms approach human level performance.
2. Using the more appropriate measure, B, only eDN approaches human performance on the Potter set and no algorithm comes close on the Thorpe dataset (even if the human performance level is reduced by 10% to compensate for prior instruction).
3. Using the measure C tailored for the Thorpe dataset, eDN leads the pack but again, somewhat below human performance.
4. Using the measure D tailored for the Potter dataset, DeepGaze II is closest, but quite below human performance.
5. Interestingly, the CENTER algorithm works almost as well as the best algorithms and sometimes outperforms all methods.

While performing these metric tests, we also plotted the locations of all the $\mathcal{P}$'s and these scatterplots are shown in Fig 7. It is easy to notice that there was a center bias in some cases as well as issues with boundary effects. Fig 6 shows scatterplots of algorithms' first $\mathcal{P}$ location for the Potter dataset in parts A (target present) and B (target absent) and the same for the Thorpe dataset in parts C and D.

The biases are striking. The AIM algorithm clearly has a problem with image boundaries. It might be ameliorated through the use of image padding as other algorithms employ, although this is not a biologically realistic solution. ITTI, eDN, oSALICON and DeepGaze II also seem

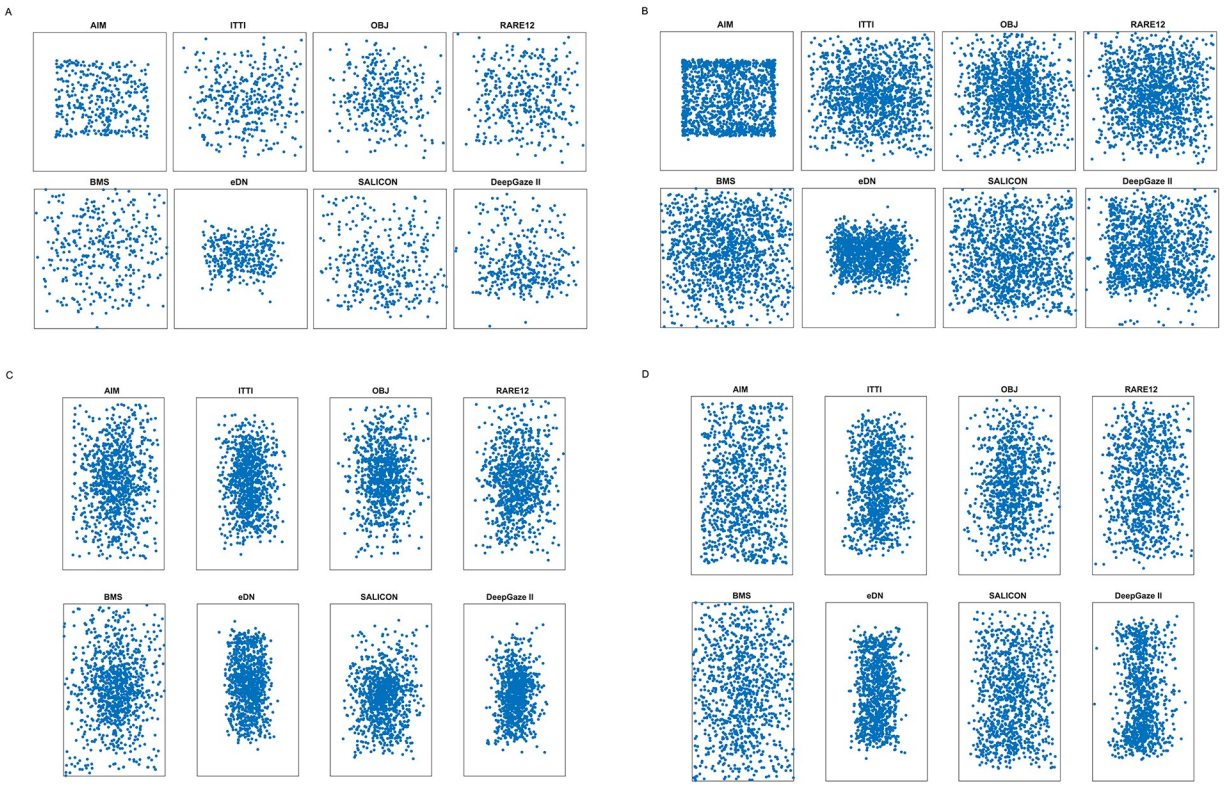

Fig 6. Plots of results using 4 performance measures. A: The percent of all first $\mathcal{P}$ that fall anywhere within the GTM for each tested algorithm and dataset. B: The percent of all first $\mathcal{P}$ that fall that fall within the GTM AND within the parafovea for each tested algorithm and dataset. C: The percent of all first $\mathcal{P}$ that fall within the GTM AND within the parafovea AND at least 27% of the GTM (by area) lies within the parafovea for each algorithm but for only the Thorpe images. D: The percent of all first $\mathcal{P}$ that fall within the GTM AND within the parafovea AND at least 41% of the GTM (by area) lies within the parafovea for each algorithm but only for the Potter images.

https://doi.org/10.1371/journal.pone.0224306.g006

to have a preference for more central $\mathcal{P}$ results. It was shown earlier that the stimulus image sets do contain center bias for their targets; however, these algorithms demonstrate a center bias for the no-target cases as well. Even though we turned off explicit center bias computations for ITTI, eDN and DeepGaze II, it seems that these algorithms have additional implicit center biases. Their good performance for the target-present cases is perhaps suspect as a result.

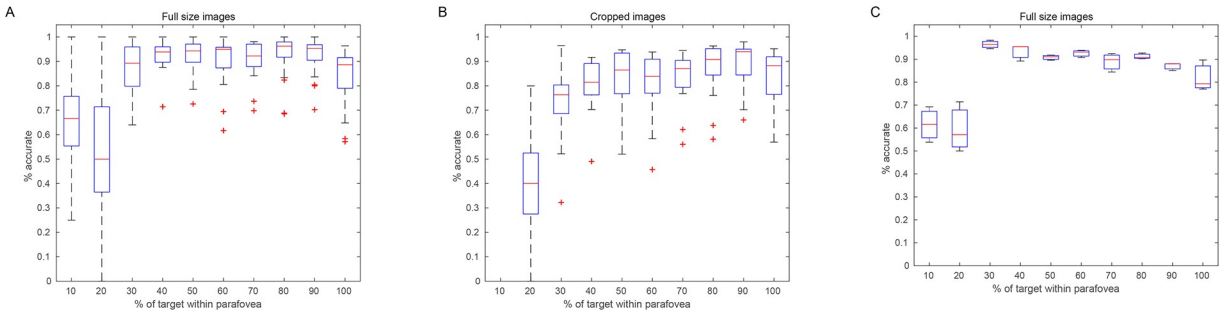

Fig 7. Scatterplots of the first attended locations $\mathcal{P}$ predicted by the saliency algorithms. A: $\mathcal{P}$ for target-present images in the Potter set. B: $\mathcal{P}$ for images with no target present from Potter set. C: $\mathcal{P}$ for target-present images in the Thorpe set. D: $\mathcal{P}$ for images with no target present from Thorpe set.

https://doi.org/10.1371/journal.pone.0224306.g007

Algorithm biological plausibility

Although many computer vision algorithms take significant inspiration from the human visual system, few embody that inspiration in a biologically realistic manner and most include extensions and enhancements in an attempt to outperform humans. It is difficult to evaluate such algorithms as to their biological plausibility but there is at least one tool that can be used. Feldman & Ballard [13] explicitly linked computational complexity to neural processes, and Thorpe & Imbert [69] further add to this as described earlier. We thus constrain biologically plausible algorithms to those requiring no more than 15 or so layers of neural computations. Since the algorithms we are testing do not deal with the full problem of categorization but only reflect the saliency computation stage, one might expect a much smaller time constraint, i.e., significantly fewer than 15 layers of computation.

Table 1 gives a coarse evaluation of the number of levels of computation, using the approximate criteria just described, for each of our tested algorithms. The AIM, BMS, ITTI, and eDN algorithms seem well within the timing constraints stated, with DeepGaze II and oSALICON outside the constraint of significantly fewer than 15 layers. Of those that do fall within the time step constraint, only eDN shows good performance, primarily on the Potter set. The depth and style of computation of the DeepGaze II and oSALICON algorithms mimics a full feedforward pass through the visual hierarchy. This perhaps argues for an incremental selection process which would be a valid possibility in human processing as well, although none of the models cited in this paper consider it (note Treisman's attenuated selection idea [25]).

Human categorization performance to parafoveal stimuli

The original stimulus set was revisited for human categorization performance to test the hypothesis that humans do not require the full image to achieve their high level of performance, and perhaps only the portion of the stimulus that is seen in an observer's parafovea was required. If the test images for rapid human categorization are center-biased as was demonstrated earlier for both the Thorpe and Potter datasets, this means there is little need for humans to require a shift in ROI if the subject's parafovea is at the image center since the target

Table 1. Number of neural processing layers in saliency algorithms. For each of the tested algorithms, an estimate of the number of neural-equivalent processing layers is presented. The AIM, BMS, ITTI, CAS, eDN algorithms seem within the timing constraints stated, with oSALICON and DeepGaze II outside.

Algorithm	Processing steps	Depth
AIM	Feature filter → Density estimation → Self-information	3
BMS	Feature split → Threshold feature channels → Connected components + Normalization → Average + Dilation	4
ITTI	DoG and Gabor Filters → Average features + Normalization → Average channels	3
eDN	Ensemble of CNNs (max 3 layers deep) → SVM combination of CNN output	4
OBJ	4 parallel streams: [SR saliency—depth1; Patch colour contrast—depth 1; Edge detection → Edge counting—depth 2; Superpixel straddling—depth 5] → Bayesian integration]	6*
RARE2012	PCA colour decomposition → log-Gabor filtering → Averaging and normalization of log-Gabor scales → Gaussian pyramid + Density estimation → Self-information of channels → Weighted average within channels → Weighted average between channels	7
oSALICON	Two-stream VGG-16 fine-tuned to saliency detection	16
DeepGaze II	Extract features from VGG-19 → readout network	19

* The equivalent convolutional depth of the superpixel step is taken to be $log(n)$ convolutions (where n is the number of pixels in the image), which works out to be approximately 5 layers for the images dealt with here.

https://doi.org/10.1371/journal.pone.0224306.t001

is usually right there too. This is necessarily true since subjects are instructed to maintain a center gaze. One can thus ask whether the parafovea is a sufficient ROI so that there would be no need to adjust its position in order to obtain good performance. We asked several questions:

1. What is the relationship between accuracy of categorization and the portion of the target that falls within the parafovea?
2. If the test images are cropped to be only the portion within the parafovea (that is, a circular region with 2.5˚ radius), what is human categorization performance?
3. What is the relationship between accuracy and portion of target within the parafovea for the 3 top performing algorithms (oSALICON, eDN and DeepGaze II)?

The results are shown in Fig 8 respectively (the experimental procedure is detailed in S3 Text). The central red mark within each box indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. Fig 8C shows performance of the top 3 saliency algorithms. We used measure A to compute average accuracy on images within bins representing % of target covered by parafovea.

Our experiment roughly duplicates Thorpe's results for the full, original images using his experimental parameters and protocol, as Fig 8A shows. High performance corresponded to at least 30% of the target being present in the parafoveal region. Even when the images are cropped to blank out extra-parafoveal portions, performance remained quite high as long as 30-40% of the target was inside the parafovea as seen in Fig 8B. Algorithm performance showed the same characteristic with the full images in Fig 8C; high accuracy was obtained whenever over 30% of the target was present in the parafovea. In other words, to a good approximation, both algorithm and human performance can be predicted by % target in the parafovea; setting a ROI within the image via any method would play little or no role. In addition, these results provide justification for our definition of performance measures C and D given earlier; for the Thorpe images, we had defined a positive hit if $\mathcal{P}$ falls within both the GTM AND the parafovea AND at least 27% of the GTM (by area) lies within the parafovea. From the results in Fig 8A, a figure perhaps closer to 30-35% might have been better. The definition we used was thus generous in favor of the algorithms.

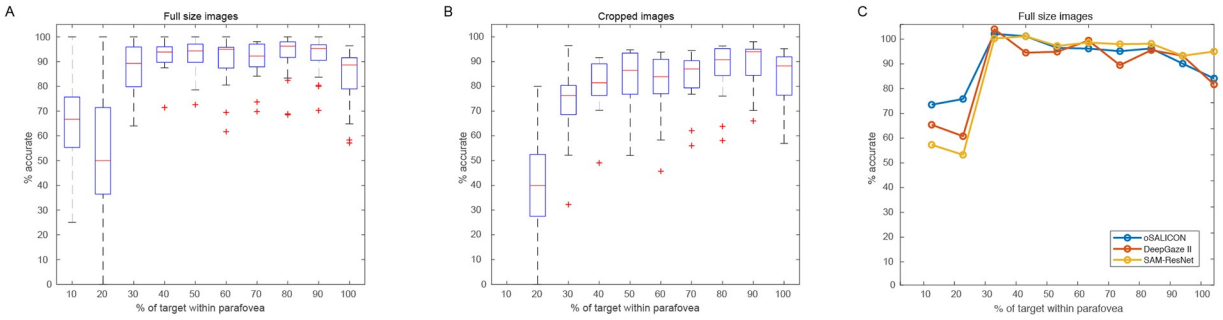

Fig 8. Results of human and algorithmic categorization performance. A,B: Box plots of human categorization performance plotted as a function of the percentage of the target within the parafovea. In the "Full size" condition (A) we used the full, original images in Thorpe's dataset. In "Cropped" condition (B) images from Thorpe's dataset were cropped so that the image portion outside the parafovea is set to a light grey. Recall that Thorpe reported 94% accuracy across the full image set. C: Performance of the top 3 saliency algorithms (oSALICON, eDN and DeepGaze II). The plot shows the percentage of correct responses (vertical axis) vs the percentage of the target within the parafovea.

https://doi.org/10.1371/journal.pone.0224306.g008

Discussion

The analysis just presented is relevant only to the role saliency computation might play for human vision. It does not directly address the role of saliency in computer vision; however, it does highlight the fact that the use of saliency computation in computer vision might have no biological motivation or justification. In machine vision, it might well play an important role in limiting the extent of the image that needs to be processed by selecting more relevant image portions (the arguments from computational complexity described in [7], [17], [28], [70] conclude exactly this point). Within the set of saliency algorithms examined, we also uncovered some interesting, and not previously revealed, biases that certainly affect their overall performance. The ITTI, eDN, oSALICON and DeepGaze II algorithms seem to have a preference for more central $\mathcal{P}$ results (even after the built-in center bias was turned off for ITTI, eDN and DeepGaze II, while oSALICON does not include an explicit center prior). This places some doubt on their strong performance for the target-present images. The rest of this section will focus on several specific points of the analysis. The overall conclusion is that neither classic nor modern saliency algorithms, in their supporting role of ROI prediction, would lead to the same high level of categorization performance in humans. Further, a closer analysis of human performance shows that there is little point to an accurate feedforward point or region of interest prediction in the original experiments.

Firstly, it is important to explain why it is justified to test existing saliency algorithms in this manner when none of them can accept nor use task specifications or prior instruction. The only Potter results we use as comparison are those without subjects receiving prior guidance, while those we use of Thorpe included uniform guidance for expected category. It has already been stated that in the Potter et al. experiment [12] without prior expectation, accuracy was roughly 10% lower than with prior guidance. van der Heijden et al. [71] also reached a similar and thus consistent conclusion, finding a 12% relative change in performance between spatial cue present and absent conditions. It thus seems reasonable to assume a similar decrease in the Thorpe experiment. That is, accuracy would likely be only about 10% lower if subjects had no instruction as to category. If we reduce the Thorpe image set human performance by 10%, then in Fig 6A oSALICON and DeepGaze II exceed human performance, but the conclusions from Fig 6B and 6C remain the same. Additionally, both algorithms do not seem to fit within the computational layer constraints. Finally, in Fig 8C, algorithm performance is shown comparable to human performance for sufficient target presence within the parafovea, i.e., the saliency prediction is not needed.

There is a problem with the set of metrics we used that should be acknowledged. Following measures C or D for the elk image of Fig 3, none of the algorithms tested would give a correct hit if the predicted P falls anywhere within the red region in the figure. This is not because of any fault of the algorithm but rather the definition of the metric itself which is tied to the position and scale of the target with respect to the observer's parafovea. This will lead to some amount of under-estimation of the accuracy of the algorithms, but the goal was to estimate how much of the target should be in the parafovea in order for a human to recognize it at the observed levels. It has already been acknowledged that the measure was not completely accurate. On the other hand, one can imagine an image where the target is small and completely within the parafovea, where a saliency algorithm hits it directly, but is not recognized by the observer because it is too small or may lie within distracting background elements. This would over-estimate accuracy with respect to humans. These two tendencies may balance each other to some degree. Nevertheless, our human experiments, shown in Fig 8, inform us that the assumptions made in defining the measures C and D are reasonable.

Table 2. Comparison of human and monkey categorization performance with the Ideal Decision Stage results. The table provides a comparison of performance for the Fabre-Thorpe et al. experiment [72], our own human experiments, and the Ideal Decision Stage we have assumed. Accuracy is computed as (TP+TN)/P+N. The TP entry for the Ideal Decision Stage reflects the best performing saliency algorithm from Fig 6C. Recall that the Thorpe et al. paper [11] reported average human accuracy of 94% correct.

		Accuracy	TP	FN	TN	FP
Fabre-Thorpe et al. monkey experiments	New Images	0.84	0.99	0.01	0.69	0.31
	Familiar Images	0.89	0.96	0.04	0.83	0.31
Our human experiments	Full Images	0.93	0.94	0.06	0.93	0.07
	Cropped Images	0.85	0.88	0.12	0.85	0.15
Ideal Decision Stage	Full Images	0.85	0.70	0.30	1.00	0.00

https://doi.org/10.1371/journal.pone.0224306.t002

We return to the issue of accuracy measures used now looking at them from a signal detection perspective. The accuracy measures reported by Thorpe and colleagues represent the averaged sum of True Positives plus True Negatives (TP+TN). We were thus constrained in our comparison wishing to align our conclusions to the human performance they reported. However, a later paper from Thorpe's group does provide an opportunity for a better signal detection analysis, with the difference that the experimental subjects are rhesus monkeys rather than humans. Our own human experimental data is also amenable to this more complete analysis.

Fabre-Thorpe et al. [72] considered rapid categorization tasks of natural images by rhesus monkeys. They note that the task presented to their subjects used the same stimulus types as the Thorpe et al. work [11], with similar methods (including additional direct human tests), and they observed similar results, leading them to conclude that humans and monkeys likely use very similar processes for these tasks. The fact that the two sets of experiments were preformed in the same lab adds credibility to their assertion. It is therefore reasonable to compare our results to this paper. In contrast to the Thorpe et al. paper [11] where only accuracy is reported, this later paper provides a fuller report of performance. The first group of rows of Table 2 gives the results from Fabre-Thorpe et al. [72] while the middle group of rows give our experiments (described earlier).

For our saliency computations each algorithm produced a point of interest regardless of image content so it is not possible to make a direct comparison. Let us assume an Ideal Decision Stage (IDS). This stage receives the fixation point from a saliency algorithm, knows what the target is, and then always outputs the correct conclusion for that point. If the fixation prediction lies within the target object (as defined by our measure C), then the output of IDS is always 'yes'. For the target-absent cases, the output will always be 'no'; it does not matter where the fixation point is, it never points to a target. In other words, for target-present trials, the output will be 'yes', a True Positive if the P is close enough to the target centroid (measure C). If P is not close enough, then the IDS will yield a 'no', a False Negative. For target-absent trials, the output of the IDS will be 'no', thus a True Negative, so TN = 1.00. There is no possibility of a False Positive since this is an ideal decision, so FP = 0 always. These are entered in the final row of Table 2 which provides the performance of the IDS. It should be clear that even with the Ideal Decision Stage assumption, saliency algorithms do not approach human nor monkey performance.

From the saliency algorithm point of view, the algorithm believes it has a potential target for all trials and does not discriminate between target present or absent scenarios. The ideal decision stage, of course, does not know which scenario is being presented. However, since it processes only the point/region of interest, it necessarily would make errors for target-present cases where the saliency algorithms produce poor predictions but will always be correct for

target-absent cases, because there could be no match to a target even though it would not have verified the absence of the target by examining the whole image. As our tests show, existing saliency algorithms do not reach the level of human accuracy and as a result, the detector's upper bound for True Positives would be the same as that of the saliency algorithms (whose best correct performance for target-present trials is 70% according to Fig 6C; thus TP = 0.70 and FN = 0.30 in the table). The saliency algorithms always produce a fixation prediction; however, in this categorization context this means that they always produce a misleading prediction for the ideal decision stage for target-absent trials. Since the whole process occurs within 150ms for both 'yes' or 'no' output, and stimulus exposure is so short, there is no time for additional processing. That is, there is no time to test that prediction, and once it is confirmed that it does not include a target, to try again until the full image is checked. Thus the Ideal Decision Stage will always be correct: the stimulus has no target. Saliency suggests a predicted fixation, it is checked and rejected. But it would be for the wrong reason. Since only one ROI is checked the overall system cannot be certain. Our assumption for the Ideal Decision Stage applies only to that stage not to the whole system. Note that human performance on target-absent cases is not perfectly accurate (see Table 2) so it seems that humans do not always correctly check the entire image either.

It is a reasonable to consider what would be appropriate for the output of a saliency algorithm for the target-absent stimuli. Firstly, it seems important for the algorithm to include knowledge of what the target is in order for it to able to distinguish targets from non-targets. The classic saliency definition has no such component; it is purely a feedforward computation depending on local image contrast alone. Even if the definition changed to include such a top-down component, an output based on a maximal local contrast computation would not suffice. For a target-absent stimulus, some separate, more global, computation seems required, perhaps of the type argued by Herzog & Clarke [73]. Could it suffice to use a variance detector that gives a global measure of low variability across an image if there is no target? This is unlikely without knowledge of the target affecting the determination. In any case, this alone could not tell the difference between an image with only regions of low interest (i.e., low local contrast) and an image with many salient regions. It would seem that some absolute measure of saliency is needed rather than a relative one (something which is impossible to do when saliency maps normalize their output as standard practice). These characteristics no longer come close to the definition nor practice of saliency computations as seen between 1985 and the present. They may however, point to directions for future computational as well as human experimental work. This potential notwithstanding, the work reported here means that it is highly unlikely that a strictly feedforward and spatially local process—as the early selection concept dictates—can suffice to drive human rapid visual categorization.

If early selection guides the process in humans, and since there is no time to check more than one ROI, then one might think that there might be some other way for checking the whole image. Perhaps humans use an entirely separate parallel stream that not only takes at most 150 ms, but considers the full image (see the 2nd paragraph of the discussion in the 1998 Fabre-Thorpe et al. paper [72] where they argue against a sequential component to this task). The second parallel process needs to be "on" always—there is no controller to decide if the saliency stage output is valid or not before deploying it—there is no time for this. So if it is always on, then a decision stage is needed to decide which output is the one to report—the one guided by saliency or the one not so guided. Since the one not guided makes a global determination, and the one guided by saliency only a local one, the global one should always be preferred, if the system is a rational agent. But all this brings us back to the original hypothesis—a single feedforward pass that takes 150ms and has the full image as input where saliency plays no role.

It is certainly true that a better saliency algorithm that fits within the strategy first outlined by Koch & Ullman [18] may yet be discovered. A still different possibility could be that even within the parafovea, some kind of early selection is taking place perhaps performing a tentative figure-ground segmentation and that selected figure is then passed on for further processing. But this seems to be simply changing the scale of the image; early selection within the fovea still has the same problem—how to be certain that a target is not elsewhere within the parafovea, and thus, this possibility does not suffice to solve the problem. With either possibility, the fact that the target-absent images would remain incompletely processed remains.

Conclusions

The widely held position that visual saliency computation occurs early during the feedforward visual categorization process in human vision was tested and we found no support for it. It is emphasized that the conclusion applies only to human vision and not any saliency role useful for machine vision systems. In fact, our experiments have shown that many machine algorithms, freed from anatomical or resource constraints that bind the human visual system, perform very well.

This is not to say that saliency computation has no role in any other aspect of human vision. In Tsotsos et al. [74] we describe a novel eye fixation prediction algorithm that employs several forms of saliency computation but not as selection for categorization tasks. It is a hybrid model that combines the positive elements of early selection, late selection, and more. We provide arguments that a cluster of conspicuity representations drives eye fixation selection, modulated by task goals and fixation history. Quantitative evaluation of this proposal shows performance that falls within the limits of human performance evaluation, and is far superior to any of the saliency methods tested [75]. Thus, visual saliency has at least the important role of participating in eye fixation computations.

Our experiments have shown that no tested algorithm can provide a sufficiently accurate first region-of-interest prediction to drive categorization results at human behaviour levels. In fact, little is gained by all the effort in comparison to the CENTER control model we tested. The many models and theories of human visual information processing, although inspiring and useful for many years of research, have served their role as important stepping stones on the path to understanding vision, but now may need to be reconsidered. Those saliency algorithms which do approach human performance seem too computationally expensive to also be biologically plausible as early selection mechanisms. It should be noted that the computational expense is not so large as to make them completely implausible; they could point to a continuous or incremental selection mechanism (as opposed to early or late) and this might be an interesting direction for future exploration. However, there is no provision in any algorithm for the target-absent stimuli. They cannot provide a 'no target' result in the same processing time; the salience-based processing strategy forces a serial search without the global computation a correct target-absent conclusion requires. We also tested human visual categorization and found that human performance seems to not need early salience. It appears sufficient for good categorization that some reasonable amount of the target appears in a subject's parafovea. Human performance was strongly predicted simply by the spatial relationship between target and the observer's parafovea, leading to the conclusion that a region-of-interest derived by any means adds little to human performance for conditions where gaze is fixed on the image center.

Supporting information

S1 Fig. Sample images from the Thorpe dataset with and without targets (animals). The top row shows full images used in the first experimental condition and the bottom row shows

the same images cropped to the size of the parafovea, used in the second experimental condition.
(TIF)

S1 File. Experiment data. Mean accuracy values for 17 human subjects and 3 saliency algorithms. Values are aggregated over different percentages of object within the parafovea region (between 10% and 100% of target by area).
(ZIP)

S1 Text. Saliency models. Links to the publicly available implementations of saliency algorithms used in this paper.
(PDF)

S2 Text. Human retina characteristics. Justification of our focus on parafoveal region in the human experiment and analysis of saliency algorithms.
(PDF)

S3 Text. Experimental details for the parafoveal stimuli test. Description of the human experiment and discussion of results.
(PDF)

Acknowledgments

The authors are grateful for the datasets and advice provided by Simon Thorpe, Molly Potter, and Brad Wyble. We thank Sang-Ah Yoo for assistance with designing the human experiment.

Author Contributions

Conceptualization: John K. Tsotsos, Iuliia Kotseruba, Calden Wloka.

Data curation: Iuliia Kotseruba.

Formal analysis: Iuliia Kotseruba, Calden Wloka.

Funding acquisition: John K. Tsotsos.

Investigation: John K. Tsotsos, Iuliia Kotseruba, Calden Wloka.

Methodology: John K. Tsotsos, Iuliia Kotseruba.

Project administration: John K. Tsotsos.

Resources: John K. Tsotsos.

Software: Iuliia Kotseruba.

Supervision: John K. Tsotsos.

Validation: Iuliia Kotseruba.

Visualization: Iuliia Kotseruba.

Writing – original draft: John K. Tsotsos.

Writing – review & editing: John K. Tsotsos, Iuliia Kotseruba, Calden Wloka.

References

1. Rosenblatt F. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. Washington DC: Spartan; 1965.

2. Fukushima K. Cognitron: A self-organizing multilayered neural network. Biological cybernetics. 1975; 20(3-4):121–136. https://doi.org/10.1007/bf00342633 PMID: 1203338
3. Fukushima K, Miyake S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Competition and cooperation in neural nets. Springer; 1982. p. 267–285.
4. Rumelhart DE, McClelland JL. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge MA: MIT Press; 1986.
5. LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. In: Arbib MA, editor. The handbook of brain theory and neural networks. Cambridge MA: MIT Press; 1995. p. 255–258.
6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
7. Tsotsos JK. A 'complexity level' analysis of immediate vision. International Journal of Computer Vision. 1988; 1(4):303–320. https://doi.org/10.1007/BF00133569
8. Potter MC, Levy EI. Recognition memory for a rapid sequence of pictures. Journal of experimental psychology. 1969; 81(1):10–15. https://doi.org/10.1037/h0027470 PMID: 5812164
9. Potter MC, Faulconer BA. Time to understand pictures and words. Nature. 1975; 253(5491):437–438. https://doi.org/10.1038/253437a0 PMID: 1110787
10. Potter MC. Meaning in visual search. Science. 1975; 187(4180):965–966.
11. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. nature. 1996; 381 (6582):520–522. https://doi.org/10.1038/381520a0 PMID: 8632824
12. Potter MC, Wyble B, Hagmann CE, McCourt ES. Detecting meaning in RSVP at 13 ms per picture. Attention, Perception, & Psychophysics. 2014; 76(2):270–279. https://doi.org/10.3758/s13414-013-0605-z
13. Feldman JA, Ballard DH. Connectionist models and their properties. Cognitive science. 1982; 6(3):205–254. https://doi.org/10.1207/s15516709cog0603_1
14. Fukushima K. A neural network model for selective attention in visual pattern recognition. Biological Cybernetics. 1986; 55(1):5–15. https://doi.org/10.1007/bf00363973 PMID: 3801530
15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997; 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 PMID: 9377276
16. Sutskever I. Training Recurrent Neural Networks [PhD Thesis]. University of Toronto; 2012.
17. Tsotsos JK. The complexity of perceptual search tasks. In: Proceedings of 11th International Joint Conference on Artificial Intelligence. vol. 89; 1989. p. 1571–1577.
18. Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence. Springer; 1987. p. 115–141.
19. Treisman AM, Gelade G. A feature-integration theory of attention. Cognitive psychology. 1980; 12(1):97–136. https://doi.org/10.1016/0010-0285(80)90005-5 PMID: 7351125
20. Broadbent D. Perception and communication. Pergamon Press, NY; 1958.
21. Deutsch JA, Deutsch D. Attention: Some theoretical considerations. Psychological review. 1963; 70(1):80–90. https://doi.org/10.1037/h0039515 PMID: 14027390
22. Mackay DG. Aspects of the theory of comprehension, memory and attention. Quarterly Journal of Experimental Psychology. 1973; 25(1):22–40. https://doi.org/10.1080/14640747308400320
23. Moray N. Attention: Selective processes in vision and hearing. London, Hutchinson Educational; 1969.
24. Norman DA. Toward a theory of memory and attention. Psychological review. 1968; 75(6):522–536. https://doi.org/10.1037/h0026699
25. Treisman AM. The effect of irrelevant material on the efficiency of selective listening. The American Journal of Psychology. 1964; 77(4):533–546. https://doi.org/10.2307/1420765 PMID: 14251963
26. Clark JJ, Ferrier NJ. Modal control of an attentive vision system. In: Proceedings of the Second IEEE International Conference on Computer Vision; 1988. p. 514–523.
27. Sandon PA. Simulating visual attention. Journal of Cognitive Neuroscience. 1990; 2(3):213–231. https://doi.org/10.1162/jocn.1990.2.3.213 PMID: 23972045
28. Culhane SM, Tsotsos JK. An attentional prototype for early vision. In: Proceedings of the European Conference on Computer Vision; 1992. p. 551–560.
29. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(11):1254–1259. https://doi.org/10.1109/34.730558

30. Bylinskii Z, DeGennaro EM, Rajalingham R, Ruda H, Zhang J, Tsotsos JK. Towards the quantitative evaluation of visual attention models. Vision research. 2015; 116:258–268. https://doi.org/10.1016/j.visres.2015.04.007 PMID: 25951756

31. Bruce ND, Wloka C, Frosst N, Rahman S, Tsotsos JK. On computational modeling of visual saliency: Examining what's right, and what's left. Vision research. 2015; 116:95–112. https://doi.org/10.1016/j.visres.2015.01.010 PMID: 25666489

32. Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F. What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;. https://doi.org/10.1109/TPAMI.2018.2815601 PMID: 29993800

33. Tsotsos JK, Eckstein MP, Landy MS. Computational models of visual attention. Vision research. 2015; 116(Pt B):93. https://doi.org/10.1016/j.visres.2015.09.007 PMID: 26420739

34. Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention. arXiv preprint arXiv:14127755. 2014;.

35. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. International Journal of Computer Vision. 2018; 126(10):1084–1102. https://doi.org/10.1007/s11263-017-1059-x

36. Shashua A, Ullman S. Structural Saliency: The Detection Of Globally Salient Structures using A Locally Connected Network. In: Proceedings of IEEE International Conference on Computer Vision; 1988. p. 321–327.

37. Olshausen BA, Anderson CH, Van Essen DC. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. Journal of Neuroscience. 1993; 13(11):4700–4719. https://doi.org/10.1523/JNEUROSCI.13-11-04700.1993 PMID: 8229193

38. Itti L, Koch C. Computational modelling of visual attention. Nature reviews neuroscience. 2001; 2(3):194. https://doi.org/10.1038/35058500 PMID: 11256080

39. Walther D, Itti L, Riesenhuber M, Poggio T, Koch C. Attentional selection for object recognition—a gentle way. In: International Workshop on Biologically Motivated Computer Vision; 2002. p. 472–479.

40. Li Z. A saliency map in primary visual cortex. Trends in cognitive sciences. 2002; 6(1):9–16. https://doi.org/10.1016/S1364-6613(00)01817-9 PMID: 11849610

41. Zhaoping L. Understanding vision: theory, models, and data. Oxford University Press, USA; 2014.

42. Deco G, Rolls ET. A neurodynamical cortical model of visual attention and invariant object recognition. Vision research. 2004; 44(6):621–642. https://doi.org/10.1016/j.visres.2003.09.037 PMID: 14693189

43. Itti L. Models of bottom-up attention and saliency. In: Neurobiology of attention. Elsevier; 2005. p. 576–582.

44. Chikkerur S, Serre T, Tan C, Poggio T. What and where: A Bayesian inference theory of attention. Vision research. 2010; 50(22):2233–2247. https://doi.org/10.1016/j.visres.2010.05.013 PMID: 20493206

45. Zhang Y, Meyers EM, Bichot NP, Serre T, Poggio TA, Desimone R. Object decoding with attention in inferior temporal cortex. Proceedings of the National Academy of Sciences. 2011; 108(21):8850–8855. https://doi.org/10.1073/pnas.1100999108

46. Buschman TJ, Kastner S. From behavior to neural dynamics: an integrated theory of attention. Neuron. 2015; 88(1):127–144. https://doi.org/10.1016/j.neuron.2015.09.017 PMID: 26447577

47. Yan Y, Zhaoping L, Li W. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. Proceedings of the National Academy of Sciences. 2018; 115(41):10499–10504. https://doi.org/10.1073/pnas.1803854115

48. Horwitz GD, Newsome WT. Separate signals for target selection and movement specification in the superior colliculus. Science. 1999; 284(5417):1158–1161. https://doi.org/10.1126/science.284.5417.1158 PMID: 10325224

49. Kustov AA, Robinson DL. Shared neural control of attentional shifts and eye movements. Nature. 1996; 384(6604):74. https://doi.org/10.1038/384074a0 PMID: 8900281

50. McPeek RM, Keller EL. Saccade target selection in the superior colliculus during a visual search task. Journal of neurophysiology. 2002; 88(4):2019–2034. https://doi.org/10.1152/jn.2002.88.4.2019 PMID: 12364525

51. Koch C. A theoretical analysis of the electrical properties of an X-cell in the Cat's LGN: Does the spine-triad circuit subserve selective visual attention. Artificial Intelligence Memo. 1984; 787.

52. Sherman S, Koch C. The control of retinogeniculate transmission in the mammalian lateral geniculate nucleus. Experimental Brain Research. 1986; 63(1):1–20. https://doi.org/10.1007/bf00235642 PMID: 3015651

53. Petersen SE, Robinson DL, Morris JD. Contributions of the pulvinar to visual spatial attention. Neuropsychologia. 1987; 25(1):97–105. https://doi.org/10.1016/0028-3932(87)90046-7 PMID: 3574654

54. Posner MI, Petersen SE. The attention system of the human brain. Annual review of neuroscience. 1990; 13(1):25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325 PMID: 2183676

55. Robinson DL, Petersen SE. The pulvinar and visual salience. Trends in Neurosciences. 1992; 15(4):127–132. https://doi.org/10.1016/0166-2236(92)90354-b PMID: 1374970

56. Thompson KG, Bichot NP, Schall JD. Dissociation of visual discrimination from saccade programming in macaque frontal eye field. Journal of neurophysiology. 1997; 77(2):1046–1050. https://doi.org/10.1152/jn.1997.77.2.1046 PMID: 9065870

57. Gottlieb JP, Kusunoki M, Goldberg ME. The representation of visual salience in monkey parietal cortex. Nature. 1998; 391(6666):481. https://doi.org/10.1038/35135 PMID: 9461214

58. Bruce ND, Tsotsos JK. Saliency, attention, and visual search: An information theoretic approach. Journal of vision. 2009; 9(3):5–5. https://doi.org/10.1167/9.3.5 PMID: 19757944

59. Bylinskii Z, Judd T, Borji A, Itti L, Durand F, Oliva A, et al. MIT saliency benchmark; 2015.

60. Huang X, Shen C, Boix X, Zhao Q. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 262–270.

61. Zhang J, Sclaroff S. Saliency detection: A boolean map approach. In: Proceedings of the IEEE international conference on computer vision; 2013. p. 153–160.

62. Vig E, Dorr M, Cox D. Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 2798–2805.

63. Riche N, Mancas M, Duvinage M, Mibulumukini M, Gosselin B, Dutoit T. Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. Signal Processing: Image Communication. 2013; 28(6):642–658.

64. Kümmerer M, Wallis TS, Bethge M. DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:161001563. 2016;.

65. Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. In: Proceedings of the IEEE International Conference on Computer Vision; 2009. p. 2106–2113.

66. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.

67. Alexe B, Deselaers T, Ferrari V. What is an object? In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on; 2010. p. 73–80.

68. Borji A, Cheng MM, Jiang H, Li J. Salient object detection: A benchmark. IEEE transactions on image processing. 2015; 24(12):5706–5722. https://doi.org/10.1109/TIP.2015.2487833 PMID: 26452281

69. Thorpe SJ, Imbert M. Biological constraints on connectionist modelling. Connectionism in perspective. 1989; p. 63–92.

70. Tsotsos JK, Culhane SM, Wai WYK, Lai Y, Davis N, Nuflo F. Modeling visual attention via selective tuning. Artificial Intelligence. 1995; 78(1-2):507–545. https://doi.org/10.1016/0004-3702(95)00025-9

71. van der Heijden AH, Schreuder R, Wolters G. Enhancing single-item recognition accuracy by cueing spatial locations in vision. The Quarterly Journal of Experimental Psychology Section A. 1985; 37 (3):427–434. https://doi.org/10.1080/14640748508400943

72. Fabre-Thorpe M, Richard G, Thorpe SJ. Rapid categorization of natural images by rhesus monkeys. Neuroreport. 1998; 9(2):303–308. https://doi.org/10.1097/00001756-199801260-00023 PMID: 9507973

73. Herzog MH, Clarke AM. Why vision is not both hierarchical and feedforward. Frontiers in computational neuroscience. 2014; 8:135. https://doi.org/10.3389/fncom.2014.00135 PMID: 25374535

74. Tsotsos J, Kotseruba I, Wloka C. A focus on selection for fixation. Journal of Eye Movement Research. 2016; 9(5).

75. Wloka C, Kotseruba I, Tsotsos JK. Active fixation control to predict saccade sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 3184–3193.