What am I searching for?

Mengni Zhang 1,2,4, Jiashi Feng 1, Joo Hwee Lim 2, Qi Zhao 3, and Gabriel Kreiman 4

to submit to IROS 2019

Address correspondences to gabriel.kreiman@tch.harvard.edu

1 National University of Singapore, Singapore
2 Institute for Infocomm Research, Astar, Singapore
3 University of Minnesota, USA
4 Harvard Medical School, Boston Children’s Hospital, USA

Abstract

Can we infer intentions and goals from a person’s actions? As an example of this family of problems, we consider here whether it is possible to decipher what a person is searching for by decoding their eye movement behavior. We conducted two human psychophysics experiments on object arrays and natural images where we monitored subjects’ eye movements while they were looking for a target object. We defined the fixations falling on the non-target objects before the target was found as “error fixations”. Using as input the pattern of “error” fixations, we developed a model (InferNet) whose goal was to infer what the target was. "Error” fixations share similar features with the sought target. The InferNet model uses a pre-trained 2D convolutional architecture to extract features from the error fixations and computes a 2D similarity map between the error fixation and all locations across the search image by modulating the search image via convolution across layers. InferNet consolidates the modulated response maps across layers via max pooling to keep track of the sub-patterns highly similar to features at error fixations and integrates these maps across all error fixations. InferNet successfully identifies the subject’s goal and outperforms all the competitive null models, even without any object-specific training on the inference task.

Introduction

Eye movements reflect rich information about the complex cognitive states of the brain, including thought processes and goals (Castelhano, Mack, and Henderson 2009; Busswell 1935; Oliva et al. 2003; Henderson et al. 2013; Betz et al. 2010; Borji and Itti 2014; Iqbal and Bailey 2004; Yarbus 1967). Additionally, with advanced eye-tracking technologies, it is now possible to monitor eye movements at high spatial and temporal resolution while controlling the task and visual environment. Therefore, eye movements provide a suitable arena to investigate how to infer a person’s goals from their actions.

Our work addresses the challenging problem of inferring what the subject is looking for in the context of a visual search task by decoding their error fixations. We define “error” fixations as the non-target fixations before the target was found. Given these error fixations, the goal is to decode what the target is (Figure 1). Several studies have shown that the error fixations during visual search are not random; those fixations are more likely to be on objects and locations that are similar to the target (Eckstein et al. 2007; Alexander and Zelinsky 2011; Wolfe 2007).

With the advancement of eye-tracking technology in wearable devices, computational models to infer the search target from human eye movements have several important application domains, such as health care, interactive user interfaces, law enforcement and virtual reality (VR). For example, gaining information of the sought object of interest would be invaluable for VR processors to provide timely feedback to players. As another example, compared with neural decoding methods based on electrode recordings inside human brains, decoding intentions in physically-disabled patients from eye movements is less invasive, has lower cost and significantly fewer potential complications.

To the best of our knowledge, there have been few attempts to build computational models that use eye fixation information for inferring what the search target is on complex natural images including cases when the sought objects may be completely “novel” to the computational models. To tackle this challenging problem, we proposed a zero-shot deep network, the Inference Network (InferNet). InferNet applies knowledge from an object recognition task on a target inference problem without any retraining. A likelihood map is computed based on feature similarity between the sub-patterns at the error fixations and the local patterns on the search image. InferNet then updates the belief of where the target of interest is across error fixations by cumulative addition of feature similarity maps modulated at each error fixation. We designed two sets of visual search experiments with object arrays and natural images, respectively, collected human eye movement data, and evaluated InferNet on these two datasets given the human error fixations in the search tasks. InferNet could successfully decode what the target was without any prior training on the inference task.

Related Works

Transfer learning. There is extensive work on networkers that can leverage knowledge from one domain to a related task (Pan and Yang 2010). Examples of transfer learning include between-class transfer in the same task (Aytar and Zisserman 2011; Lim, Salakhutdinov, and Torralba 2011;
Tommasi, Orabona, and Caputo 2010); between task transfer, such as from classification to object detection (Ren et al. 2015; Redmon et al. 2016; Liu et al. 2016) and image classification to semantic segmentation (Long, Shelhamer, and Darrell 2015). Our work focuses on task transfer by taking a network pre-trained for image classification and applying those weights on the target inference task without any fine-tuning on this new task. To our best knowledge, deciphering human intentions based on fixation patterns with no prior category-specific training (zero-shot) has not been done before and we are the first to apply feature similarities in latent space and use transfer learning to tackle this problem.

**Target decoding from fixations.** Although information about a target is available in the fixation behavior during visual search, this does not imply that subjects are able to extract this information and use it to infer a search target (Eckstein et al. 2007; Alexander and Zelinsky 2011; Wolfe 2007). Whether humans can infer the target information from other people’s fixation behavior or not remains controversial. Some researchers have reported that it is possible to decode task information from eye movements (Borji and Itti 2014; Haji-Abolhassani and Clark 2013; Cerf et al. 2008; O’Connell and Walther 2012; Borji et al. 2013; Rajashekar, Bovik, and Cormack 2006) while others have argued against otherwise (Henderson et al. 2013; Greene, Liu, and Wolfe 2012).

The focus in the current study is not about a mechanistic understanding of human visual search but on designing a computational model capable of inferring what the subject’s target is. The fact that we can perform this task at all shows that eye movements are not random and that they convey information about the goals. There are a few studies on decoding target information in the context of visual search (Borji, Lennartz, and Pomplun 2015; Zelinsky, Peng, and Samaras 2013; Rajashekar, Bovik, and Cormack 2006), but current methods are limited in using elementary search statistics (Rajashekar, Bovik, and Cormack 2006) and handcrafted features (Borji, Lennartz, and Pomplun 2015; Zelinsky, Peng, and Samaras 2013). Moreover, existing approaches have only been tested with pre-defined object classes with constrained object set sizes. These computational models do not generalize to infer any target from arbitrary classes. In contrast, what we find remarkable here is the proof-of-principle demonstration of being able to transfer learning from visual recognition to predict human intentions without any additional training whatsoever on new object categories in natural complex scenes.

**InferNet**

We provide an overview of the model, followed by a more detailed description of our proposed zero-shot deep network (InferNet, Figure 2).

**Overview**

Error fixations share more visual feature similarities with the target than with distractors (Eckstein et al. 2007; Alexander and Zelinsky 2011; Wolfe 2007) (see also Supplementary Material for feature similarity comparison between pairs of targets and error fixations versus pairs of targets and random fixations). Thus, our model is based on the idea that the location with more feature similarities for all error fixations is more likely to be the search target location. We approximate the target inference problem in feature similarity space among targets and distractors: given $T$ error fixations with coordinates $(x_i, y_i)$ where $1 \leq i \leq T$, the task is to predict a 2D probabilistic map $M_f$ of where the search target is most likely to be (Figure 2). We take the maximum on $M_f$ as the current guess location. If the cropped area centered at the current guess location overlaps with the ground truth bounding box encompassing the whole target object, the inference is deemed successful; otherwise, after each incorrect guess, the map is updated by removing the erroneous inference location on $M_f$. 

Figure 1: Illustration of the target inference problem. Human subjects were instructed to move their eyes to search for a given target (A) in the search image (B) irrespective of changes in size, rotation angles, or other format changes. The visual search task resulted in a sequence of fixations (C, yellow circles with the arrows). The red bounding box refers to the ground truth target location in the search image (not shown in the actual experiment). In this example, the subject required 2 fixations to find the target. We defined the fixations falling on the non-target objects as “error fixations”. In the target inference task, given the error fixations recorded from the psychophysics visual search task (D, yellow circle), the model is asked to infer what target object the subject was searching for out of the remaining possible objects (E, question marks in orange color, the question marks are not shown to the computational model). In this example, there is only 1 error fixation, in general, there could be anywhere from 1 to 4 error fixations in these experiments with arrays of 6 objects.
Figure 2: Architecture of InferNet. At each error fixation $i$, InferNet takes two inputs: the object $I_{ie}$ at the error fixation and the search image $I_{is}$ with the object at the error fixation inhibited with a black mask. The model consists of a pre-trained deep convolutional network that processes the objects at the error fixations (Prior Network (orange shade)) and also processes the search image (Likelihood Network (gray shade)). The weights used to process the error fixations and the search images are identical and are pre-trained for image classification (see text). The Prior Network generates feature maps in each layer from the object at error fixations $I_{ie}$ whereas the Likelihood Network generates feature maps in each layer for the search array image $I_{is}$ via a 2D convolutional neural network. Conditioned on the Prior Network, the Likelihood Network modulates the prior response maps by convolving the error fixation representation of $I_{ie}$ with the feature maps from $I_{is}$ at multiple layers, generating feature similarity maps $M_{11}, M_{12}, ..., M_{1N}$. These feature similarity maps are then resized, normalized, and concatenated. We perform max-pooling across these maps to generate the consolidated feature similarity map $M_{if}$. This process is repeated for each error fixation $i$. The final probabilistic map $M_f$ is the sum of all the individual error fixation maps. InferNet makes a decision on where the target is possibly located based on the maximum activation on $M_f$ (red dot). An inhibition of return mechanism is applied if the target is not found at the current inferred location and the next maximum on $M_f$ is selected. The error fixations are recorded from human subjects in the visual search task (Figure 1A-C). A schematic of the human psychophysics experiment in the visual search task is shown in the dash black box on the top right.

The model is based on a pre-trained deep convolutional network that is applied to the error fixations (Prior Network (PN)) and to the search image (Likelihood Network (LN)). PN takes the cropped area $I_{ie}$ of size 28 × 28 pixels centered at error fixation $i$ as input and outputs feature maps across layers. We define $I_{is}$ as the search image which has the objects at all past error fixations $i$, ... $i$ inhibited with a black mask. LN modulates the feature maps from $I_{ie}$, generating a series of likelihood maps $(M_{i1}, M_{i2}, ..., M_{ij}, ..., M_{iN})$ across different layers where $j$ denotes the index of the $j$th likelihood map $M_{ij}$ for error fixation $i$. These maps are concatenated and max-pooled to produce the final likelihood map $M_{jf}$ for error fixation $i$ which tracks the parts of the image that are most similar between $I_{ie}$ and $I_{is}$. InferNet integrates these likelihood maps $M_{ij}$ across all $T$ error fixations via element-wise-sum by assuming all the error fixations play equally important roles in contributing to the final inference map $M_f$.

Prior Network
We used a deep feed-forward network, implemented in VGG-16 (Simonyan and Zisserman 2014), and pre-trained for image classification on the ImageNet dataset (Russakovsky et al. 2015). We show that the invariant features from VGG-16 can be directly used for target inference task without any additional training. Given $I_{ie}$ at error fixation $i$, the network weights $W$ learnt from image classification extract feature maps $\varphi_{j}^{PN}(I_{ie}, W)$ at layer $j$ (orange boxes in Figure 2).

Likelihood Network
Given $I_{is}$, LN has the same network parameters $W$ as PN and extracts the feature representation of $I_{is}$ at layer $j$, $\varphi_{j}^{LN}(I_{is}, W)$ (gray boxes in Figure 2). The weights are shared between PN and LN, and both are pre-trained for image classification, not for target inference. The weights $W$ do not depend on $I_{is}$ or $I_{ie}$. The InferNet network has
no prior training with the objects or images in this study. The locations of the error fixations in $I_{is}$ are blacked out (so that the model does not indicate that the most similar location to an error fixation is the error fixation itself). The input to PN is smaller than the input to LN, hence the output $\varphi_j^{PN}(I_{ie}, W)$ is smaller than $\varphi_j^{LN}(I_{ie}, W)$. The activity of the units in LN in response to the search image is modulated by those in PN, which contain features more similar to the visual search target than distractors.

The modulation in the activation map is achieved by convolving the representation of the error fixation with the representation of the search image at multiple scales:

$$M_{ij} = m(\varphi_j^{SN}(I_{is}, W), \varphi_j^{PN}(I_{ie}, W)) \quad (1)$$

where $m(\cdot)$ is the error fixation modulation function defined as 2D convolution with kernel $\varphi_j^{PN}(I_{ie}, W)$ on the search feature map $\varphi_j^{LN}(I_{is}, W)$ where $j$ denotes the index of the $j$th feature similarity map $M_{ij}$ for error fixation $i$.

Inspired by neurophysiological recordings during visual search and attentional modulation in visual cortex (Desimone and Duncan 1995; Gilbert and Li 2013; Miconi, Groomes, and Kreiman 2015), and with the goal of capturing target properties at multiple scales and with different features, modulation is applied across multiple layers. Intuitively, if the target object shares more similarities with the error fixations in low-level features, such as similar orientations, error fixation modulation on $M_{ij}$ may be sufficient; however, if high-level features are shared between the target and the error fixations, such as surface texture, feature similarity maps at higher levels may be required. We empirically selected $N = 7$ feature similarity maps (see details in Supplementary Material). In general, it is possible to select other layers based on specific applications, or even learn which layers to select for specific problems.

Each of these feature similarity maps is up-sampled to $224 \times 224$ pixels and the final feature similarity map is max pooled at each location $(x, y)$ on $M_{ij}$ over all the $N$ intermediate maps (Table 2 reports performance separately for each feature similarity map). The model thus keeps track of all the locations which share similar sub-patterns including both low-level and high-level feature descriptors:

$$M_{ij}(x, y) = \max_{j=1}^{N} M_{ij}(x, y) \quad (2)$$

Combination of maps and target inference

The feature similarity maps $M_{ij}$ are summed over all $T$ error fixations:

$$M_f(x, y) = \sum_{i=1}^{T} M_{ij}(x, y) \quad (3)$$

We assume all error fixations play equally important roles in inferring the search target. In general, it is possible to use a weighted summation where some error fixations are more important than the rest depending on the applications. InferNet selects the maximum of the $M_f$ map. If the cropped area centered at the current guess location overlaps with the ground truth bounding box encompassing the whole target object, the inference is deemed successful and the inference stops. Otherwise, that location is inhibited and the next maximum is selected.

Evaluation

To evaluate performance of InferNet, we computed the average number of guesses required over all the trials with different images as a function of the number $T$ of error fixations. The less number of guesses required, the more effective the inference process is. However, since the target inference difficulty varies, we report the relative performance $P_T$ defined as the average number of guesses required by the computational model $A_m(T)$ relative to the average number of guesses required by a chance model $A_c(T)$ on the same image and task (the chance model is defined below):

$$P_T(T) = \frac{A_m(T) - A_m(T)}{A_c(T)} \times 100 \quad (4)$$

If the computational model requires less number of guesses on average, $P_T(T)$ is greater than zero. The larger $P_T(T)$, the more efficient the inference process is.

Experiments

We tested InferNet on images containing object arrays and in natural images by evaluating the number of guesses required to correctly infer the sought target, $P_T(T)$. As benchmarks, we compared our model with other alternative null models. All the data (images, eye movements in visual search, source code) will be publicly shared upon publication.

Datasets

We consulted with psychologists and designed two sets of psychophysics visual search tasks based on (Miconi, Groomes, and Kreiman 2015): object arrays and natural images. Ten subjects (5 in each task) were first presented with the exemplar target followed by the search image (see Figure 2 for schematic illustration of our psychophysics experiment). The target was always present for all trials. We used an EyeLink D1000 eyetracker (SR Research, Canada) to record eye movements during the visual search tasks. In the target inference task, we filtered out those fixations on targets and only used error fixations obtained prior to subjects locating the target in each trial. The appearance of the target object in the search image was different from that in the target image.

Object Arrays We selected segmented objects without occlusion from natural images in the MSCOFCO dataset (Lin et al. 2014) from 6 categories: sheep, cattle, cats, horses, teddy bears and kites. Due to the uncontrolled and diverse nature of these stimuli, they may differ in low-level properties that could contribute to visual search performance. To minimize such contributions, we took several steps to normalize their low-level features (see Supplementary Material for details). Six objects (one per category) were uniformly arranged in a circle. There were 300 trials in total.
Inferences

7
2
8
[3]
[31x53]Relative performance (%)
40x33
−20
−15
−10
−5
0
5
10
15
20
25
Numbers of error fixations
Relative performance (%)
(a) Object arrays
(b) Natural images

Figure 4: Evaluation of model inference performance for object arrays (a) and natural images (b). Relative performance improvement for the computational model relative to the chance model as a function of the number of error fixations. The smaller the number of guesses, the better the inference algorithm is and the higher the relative performance improvement. The different colors denote different models: InferNet model (blue), bottom-up IttiKoch saliency (red), template matching (green), RanWeight (magenta), Chance (black). See Section Comparative Null Models for descriptions. Error bars are standard error of the mean for all trials.

We considered a pure bottom-up saliency model that has no information about the error fixations (Itti, Koch, and Niebur 1998).

RanWeight. Instead of using VGG16 (Simonyan and Zisserman 2014) pre-trained for image classification, we randomly picked weights \( W \) from a gaussian distribution with mean 0 and standard deviation 1000. The network was otherwise identical to InferNet. The random selection of weights was repeated 100 times.

Object arrays

Figure 3 shows examples illustrating how the model efficiently inferred the target location given only one or two fixations on object arrays. In the first example (Column 1-3, Row 1), a subject made one error fixation on the cow which looks visually similar to the sheep before finding the sheep. Given this single error fixation, InferNet determined

Natural Images To evaluate whether our model could generalize to infer the sought target in complex natural images, we collected 240 natural images from common object categories, such as animals (clownfish) and daily objects (alarm clock). In contrast to the object arrays, here the objects were immersed in natural background and clutter and the object classes were not restricted to 6 categories. None of the images in the data set were taken from ImageNet, the dataset used to train VGG16. Moreover, there were 140 images out of the selected 240 images containing target objects whose categories are not part of ImageNet. In other words, these objects are novel to InferNet. The target object as rendered in the target image differed from the one rendered in the search image in terms of size, pose and rotation.

Comparative Null Models

We compared our model with several alternative null models. In all cases, the alternative models proposed an inference map and the procedure to select a target was the same as with InferNet, including infinite inhibition-of-return (i.e. never selecting the same location twice).

Chance. We considered a model where the target location was chosen at random. For object arrays, we randomly chose one out of the remaining possible locations. For the natural images dataset, a random location was selected for each guess. This random process was repeated 20 times.

Template Matching. To evaluate whether pixel-level features of the error fixations were sufficient for guiding inference, we introduced a pixel-level template matching model where the inference map was generated by sliding the canonical target of size 28 × 28 pixels over the whole search image of size 224 × 224 pixels. Compared to the classical sliding window models in computer vision, this can be interpreted as an “attentional” sliding window.

IttiKoch. We considered a pure bottom-up saliency model that has no information about the error fixations (Itti, Koch, and Niebur 1998).
that the subject was probably looking for a sheep among all the five remaining distractors (red circle, Column 3, Row 1). In the second example (Column 1-3, Row 2), a subject made 2 error fixations before finding the target (horse). In this case, InferNet correctly determined the target at the 3rd guess (Column 3, Row 2).

InferNet showed an overall improvement of 3.8±3% with respect to the chance model over all error fixations (Figure 4a, blue). Even with a single error fixation as input data, InferNet could infer the target 6.87% faster than the chance model. That is, while random guessing would correctly land on the target within 3 guesses, InferNet only required 2.80±0.01 guesses on object arrays.

In Figure 4a, none of the null models reached the level of relative performance improvement shown by InferNet ($P < 4.6 \times 10^{-20}$, two-tailed t-test, $t = -9.2, df = 12128$) for all the numbers of error fixations except for the case of 4 error fixations where none of the models were above chance. Though we took precautions to normalize average low-level features on arrays, for goal inference, on any trial, InferNet can capitalize on shared IttiKoch features between error fixations and the target. Performance for the bottom-up saliency model (IttiKoch) is better than the chance model but still below InferNet which suggests that the target information embedded in error fixations is useful for target inference. The model with random weights (RandWeight) and the model with template matching (TempMatch) on pixel levels show minimal improvements from selecting random locations (Figure 4a), suggesting the discriminative features learnt from a hierarchical network for image classification are important for target inference.

**Target category inference**

In addition to inferring the target location, we tested InferNet on object category inference task. Out of 240 natural images, we selected 100 images where the target categories belong to ImageNet. For each subject, InferNet predicts the belief of possible target categories out of total 1000 categories by leveraging on the weights pre-trained on ImageNet and accumulates these belief across error fixations. Table 1 reports the accuracy of top $N$ most probable target categories inferred by InferNet based on the accumulated belief across error fixations. We have two observations. First, given even only one error fixation, the inference accuracy of InferNet surpasses the chance model (1/1000). As $N$ increases, the target category inference accuracy increases. Ideally, the accuracy of inferred top 1000 probable target

| Error Fixations | Top $N$ category inference accuracy % |
|-----------------|---------------------------------------|
| 1               | 6 8 11 13 17 29 38 55                |
| 2               | 4 9 14 20 23 33 46 65                |
| 3               | 3 10 16 25 28 38 51 72               |
| 4               | 0 13 20 20 30 39 54 74               |
| 5               | 3 11 23 28 34 42 56 74               |
| 6               | 0 14 23 31 37 45 54 77               |
| 7               | 1 15 25 31 37 47 53 78               |
| 8               | 4 17 28 37 40 48 57 80               |

Table 1: Our model performance of top $N$ inferred target category accuracy across error fixations (rows) where $N = 1, 2, ..., 128$ (columns) is shown.
categories should be 1 as the target always belong to at least one of the 1000 categories from ImageNet. Given 8 error fixations, InferNet is capable of inferring the target category correctly with accuracy of around 50% for top 32 most probable categories out of 1000 categories. Second, as InferNet takes more number of error fixations as inputs, the belief gets constantly updated and the inference becomes more accurate. This validates the error fixations carry important information revealing the target identity during visual search.

### Ablation study

To evaluate the contribution of different layers of InferNet, we tested each individual feature similarity map $M_j$ and their different combinations in object arrays and natural images. Table 2 shows our ablated models’ relative performance compared with the chance model using feature similarity maps ($M_j$) at different layers $j$ for $T$ error fixations. The layer number refers to the index in the VGG16 network (Simonyan and Zisserman 2014). The first row $M_1$ corresponds to our full model considering all feature similarity maps across layers whereas the other rows show the predictions using either only one feature similarity map from $M_{1:7}$ in Figure 2 or their combinations.

From Table 2, we have several observations: (1) Compared to the individual maps, target inference performance was generally more effective using the feature similarity maps $M_j$ in higher layers which implies that high-level features extracted at error fixations are more reliable for target inference. (2) We are also interested in exploring how the compositionality of feature similarity maps across layers reveals the identity of the target. InferNet takes max-pooling of $M_{ij}$ for error fixation $i$ and averages $M_{1:i}$ for all $T$ error fixations. Instead of max-pooling across layers, we also evaluated ablated models where the max-pooling across $N$ layers is replaced by averaging and vice versa. We did not observe any significant improvements in object arrays but different combination methods of feature similarity maps contribute dramatically differently in natural images. Our InferNet model outperforms the rest which suggests error fixations seem not to be guided by the overall target features as a whole (taking average across $N$ layers) but by sub-patterns of the search target (max-pooling across $N$ layers) which aligns with (Rajashekar, Bovik, and Cormack 2006).

(3) Our InferNet model treats all error fixations equally and only utilizes the visual feature information at the error fixations. In the last ablated model, we study the role of the locations and the sequence order of error fixations in target inference (see Supplementary Material). It is surprising that the experimental result seems to suggest the location and order information of error fixations do not matter much in target inference task.

Human visual search is variable both within-subject and between-subjects (Miconi, Groomes, and Kreiman 2015). We conducted additional psychophysics experiments and reported the results in Table 2 (last two rows). Humans were not able to solve the inference problem in object arrays but were better than InferNet in natural images, perhaps by using contextual cues (second last row). To investigate the between-subject variability, we created a new model using only fixations that are common across subjects. The result (last row) shows that in some (but not all) cases, InferNet can overcome the consequences of variability in human s-canpath patterns. However, in general, we need algorithms that can predict individual intentions in single trials, which is the goal for InferNet.

### Conclusion

We proposed a computational model to infer intentions from behaviors in the context of a visual search task. InferNet can determine what the sought target is, in object array images as well as in natural images, by using the prior set of non-target fixations. InferNet is based on transfer-learning in that it uses weights learnt for a different task. InferNet is a “zero-shot” architecture: there is no training with the specific objects or images that the model analyzes during the inference process. Leveraging on the idea that error fixations share feature similarities with the targets, InferNet builds an implicit relationship between the inference problem and the feature similarity problem. The experimental results show that InferNet significantly outperforms the comparative null models. There are many areas where the model could be improved. Most notably, inference could be enhanced by incorporating intuitive semantics in the real world (e.g. if the error fixations are mostly distributed on the ground, one could deduce that...
the target of interest would most likely not be the airplanes in the sky). Problem-specific training (e.g., weights for each layer, or weights for each error fixation) could also improve performance. The proof-of-principle demonstration in this study provides a possible inference solution to effectively study provides a possible inference solution to effectively guess what the subject is searching for in complex images and suggests that computational models can make reasonable conjectures to read the subject’s mind purely based on behavioral data.

References

Alexander, R. G., and Zelinsky, G. J. 2011. Visual similarity effects in categorical search. Journal of Vision 11(8):9–9.

Aytar, Y., and Zisserman, A. 2011. Tabula rasa: Model transfer for object category detection. In Computer Vision (ICCV), 2011 IEEE International Conference on, 2252–2259. IEEE.

Betz, T.; Kietzmann, T. C.; Wilming, N.; and König, P. 2010. Investigating task-dependent top-down effects on overt visual attention. Journal of Vision 10(3):15–15.

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

Borji, A.; Tavakoli, H. R.; Shihite, D. N.; and Itti, L. 2013. Analysis of scores, datasets, and models in visual saliency prediction. In Computer Vision (ICCV), 2013 IEEE International Conference on, 921–928. IEEE.

Borji, A.; Lennartz, A.; and Pomplun, M. 2015. What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. Neurocomputing 149:788–799.

Busswell, G. 1935. How people look at pictures: A study of the psychology of perception in art.

Castelhano, M. S.; Mack, M. L.; and Henderson, J. M. 2009. Viewing task influences eye movement control during active scene perception. Journal of vision 9(3):6–6.

Cerf, M.; Harel, J.; Huth, A.; Einh¨auser, W.; and Koch, C. 2008. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. In International Workshop on Attention in Cognitive Systems, 15–26. Springer.

Desimone, R., and Duncan, J. 1995. Neural mechanisms of selective visual attention. Annual review of neuroscience 18(1):193–222.

Eckstein, M. P.; Beutter, B. R.; Pham, B. T.; Shimozaki, S. S.; and Stone, L. S. 2007. Similar neural representations of the target for saccades and perception during search. Journal of Neuroscience 27(6):1266–1270.

Gilbert, C. D., and Li, W. 2013. Top-down influences on visual processing. Nature Reviews Neuroscience 14(5):350–363.

Greene, M. R.; Liu, T.; and Wolfe, J. M. 2012. Reconsidering yarbus: A failure to predict observers task from eye movement patterns. Vision research 62:1–8.

Haji-Abolhassani, A., and Clark, J. J. 2013. A computational model for task inference in visual search. Journal of Vision 13(3):29–29.

Henderson, J. M.; Shinkareva, S. V.; Wang, J.; Luke, S. G.; and Olejarczyk, J. 2013. Predicting cognitive state from eye movements. PloS one 8(5):e64937.

Iqbal, S. T., and Bailey, B. P. 2004. Using eye gaze patterns to identify user tasks. In The Grace Hopper Celebration of Women in Computing, 5–10.

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

Lim, J. J.; Salakhutdinov, R. R.; and Torralba, A. 2011. Transfer learning by borrowing examples for multiclass object detection. In Advances in neural information processing systems, 118–126.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In European conference on computer vision, 740–755. Springer.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In European conference on computer vision, 21–37. Springer.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3431–3440.

Miconi, T.; Grommes, L.; and Kreiman, G. 2015. There’s waldo! a normalization model of visual search predicts single-trial human fixations in an object search task. Cerebral Cortex 26(7):3064–3082.

O’Connell, T., and Walther, D. 2012. Fixation patterns predict scene category. Journal of Vision 12(9):801–801.

Oliva, A.; Torralba, A.; Castelhano, M. S.; and Henderson, J. M. 2003. Top-down control of visual attention in object detection. In Image processing, 2003. icip 2003. proceedings. 2003 international conference on, volume 1, I–253. IEEE.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10):1345–1359.

Rajashekar, U.; Bovik, A. C.; and Cormack, L. K. 2006. Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. Journal of Vision 6(4):7–7.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 91–99.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115(3):211–252.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Tommasi, T.; Orabona, F.; and Caputo, B. 2010. Safety in numbers: Learning categories from few examples with multi model networks for large-scale image recognition. In European conference on computer vision, 3081–3088. IEEE.

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

Yarbus, A. 1967. Eye movements and vision. 1967.

Zelinsky, G. J.; Peng, Y.; and Samaras, D. 2013. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. Journal of vision 13(14):10–10.