Benchmarking Gaze Prediction for Categorical Visual Search

Gregory Zelinsky Zhibo Yang Lihan Huang Yupei Chen Seoyoung Ahn
Zijun Wei Hossein Adeli Dimitris Samaras Minh Hoai

Stony Brook University, Stony Brook, NY, 11794, USA
{gregory.zelinsky, zhibo.yang, lihan.huang, yupei.chen, seoyoung.ahn,
zijun.wei, hossein.adeli, dimitrios.samaras, minh.nguyen}@stonybrook.edu

Abstract

The prediction of human shifts of attention is a widely-studied question in both behavioral and computer vision, especially in the context of a free viewing task. However, search behavior, where the fixation scanpaths are highly dependent on the viewer's goals, has received far less attention, even though visual search constitutes much of a person's everyday behavior. One reason for this is the absence of real-world image datasets on which search models can be trained. In this paper we present a carefully created dataset for two target categories, microwaves and clocks, curated from the COCO2014 dataset. A total of 2183 images were presented to multiple participants, who were tasked to search for one of the two categories. This yields a total of 16184 validated fixations used for training, making our microwave-clock dataset currently one of the largest datasets of eye fixations in categorical search. We also present a 40-image testing dataset, where images depict both a microwave and a clock target. Distinct fixation patterns emerged depending on whether participants searched for a microwave (n=30) or a clock (n=30) in the same images, meaning that models need to predict different search scanpaths from the same pixel inputs. We report the results of several state-of-the-art deep network models that were trained and evaluated on these datasets. Collectively, these datasets and our protocol for evaluation provide what we hope will be a useful test-bed for the development of new methods for predicting category-specific visual search behavior.

1. Introduction

How humans allocate their spatial attention, overtly measured by changes in eye gaze, is a question having clear benefits to both computer and behavioral vision. For behavioral vision, a model that can predict the sequence of gaze fixations made in response to an image would be a source of innumerable hypotheses for behavioral testing that would accelerate our understanding of human attention. For computer vision, an ability to predict fixation locations would similarly drive the development of next-generation systems that could intelligently anticipate a user's needs or desires. In one sense this mutual benefit has already been realized. The goal of predicting fixation behavior in a free-viewing task has fueled the development of saliency models, so much so that there is an active competition for best performance[1], and these methods are increasingly being used in computer vision applications ranging from object detection [27] to intelligent image editing/re-targeting [10]

Figure 1: A viewer's scanpath while searching for a microwave. Note the clear preference for fixating locations on the kitchen countertop, a behavior that would not be reliably captured by saliency maps from bottom-up models. Can computational models predict the priority maps underlying such search scanpaths? In this paper, we propose a behavioral dataset that provides a useful test-bed for qualitative and quantitative evaluation of this important task.

However, it is important to realize the distinction be-

[1] http://saliency.mit.edu/results_mit300.html

tween "saliency" and "priority". Priority, as the term is used in the fixation prediction literature [38], refers to a general prioritization of image locations for the purpose of predicting gaze, with the term "saliency" referring to a specific type of prioritization–one based on information solely in the image input (e.g., feature contrast, as in the saliency model by Itti et al. [15]). For this reason, saliency models are often described as "bottom up"; they will produce the same output for an image regardless of the goals of the person. Bottom-up saliency models have historically been contrasted with "top-down" models of attention control, although this dichotomy has become strained both in theory [29] and by recent saliency models explicitly or implicitly incorporating limited top-down information in their predictions [17, 18]. In general, "top-down" models of attention control recognize that the vast majority of meaningful gaze behavior is made in the service of specific tasks and goals. Moreover, these tasks or goals can be entirely arbitrary. If a person walks into an unfamiliar kitchen with the task of warming a cup of tea, their goal might be to find a microwave oven. To mediate this goal, priority should therefore be assigned to the locations in the kitchen input having features offering the most information about microwaves. But if this person's task was to check the time, their goal would be to find a clock and these features should be prioritized in the input instead. These different microwave and clock prioritizations would both be considered top-down, and different from bottom-up prioritization in that the same visual input would lead to potentially very different fixation behavior. Perhaps more useful than a bottom-up/top-down dichotomy would be to consider a set of possible priority maps equalling a set of tasks that might be engaged given an image input, with the prioritization output by a saliency model being specific to the relatively minimal task of free-viewing.

The present study focuses on goal-directed behavior, and specifically on a visual search task. Visual search, the human analog to object detection in computer vision, is arguably the simplest of goal-directed behaviors—there is an object goal, called the "target", and the task is to determine the location of this target goal in an image (or to conclude that it is absent). This goal-specific prioritization is measured behaviorally by an increase in the probability of fixating image locations having target features, with this preferential direction of attention referred to as "target guidance". Target guidance during search was first quantified using very simple targets having simple features that were known to the searcher (e.g., [34]). This work was followed by computational models that used more complex images as inputs, but still assumed perfect knowledge of the target's features (e.g., [37]). Most recently, search guidance has been shown to targets defined only by their object class [21, 22, 28, 31, 32, 35, 36, 39, 40], making the question more aligned with efforts in computer vision. The study presented in this paper used a categorical search task, and specifically asked people to search for either microwave or clock categorical targets in kitchen scenes, as in the example shown in Figure 1.

A model's success in predicting fixation behavior depends on the availability of data that can be used for training. Saliency models are again an excellent example of this, with the currently best performing models all being deep neural networks trained on the fixations of people viewing large image datasets [2, 16, 17, 33]. Here we attempt to do the same for visual search. The prediction of fixation behavior during categorical search is currently limited by the availability of training data. The few datasets that could be used to train a search model are either relatively small and limited to people [8] or large and including more target categories (six classes of animals) but from a task in which participants were instructed to "find all animals in the scene" [11] rather than a more standard search task having target-present and target-absent trials. There is also the POET dataset [24], which contains fixation data from 28 people viewing 6270 images from VOC2012 [9] depicting ten target classes (cat, dog, boat, aeroplane, horse, cow, bicycle, motorbike, sofa and dining table), but the task was two-alternative forced-choice object discrimination and not visual search. Our microwave-clock-search dataset (MCS) was collected using a categorical visual search task with interleaved target-present and target-absent trials. It contains high-quality fixations obtained under well controlled laboratory conditions, and is large enough to train deep network models. Additional effort was expended in collecting a test set of images that depicted both microwaves and clocks, and in the use of these images to evaluate the success of state-of-the-art deep learning models in predicting search scanpaths for these two target categories.

2. Behavioral Methods and Data

Two behavioral data collection efforts were conducted for this study. The first involved collecting eye movement data from people searching for targets from either the microwave or clock categories (not both). This was done for a large number of images with the goal being to create a dataset of fixation-labeled images large enough for model training. The second effort again involved collecting gaze fixation data for the same target categories, but this time for a smaller and more controlled dataset and using a larger number of participants. Here our goal was to obtain a valid ground truth for search behavior against which models could be evaluated. Both the training and testing images were selected from COCO2014 [20]. The training and testing datasets are available at `https://www3.cs.stonybrook.edu/~cvl/projects/coco_search/index.html`, and

Figure 2: Representative images and scanpaths from our dataset for viewers searching for a clock. Each image was seen by multiple viewers, and their gaze scanpaths were collected. Note the clear difference in behavior relative to the scanpath in Figure 1.

some example images and scanpaths are shown in Figure 2.

2.1. Training data

Given the practical costs and limits associated with the collection of high-quality fixation behavior (200 search images $\approx$ 1 hour of a participant's time in the laboratory), our current effort was restricted to just two target categories: microwaves and clocks. The microwave category in COCO2014 [20] has 1089 images in their training set and 512 images for validation. The clock category has 3159 training images and 1704 validation images. However, several criteria were imposed on the selection of training images: (1) images were selected only from the training sets for the microwave and clock categories and images overlapping with the testing data were excluded. (2) Images were excluded if they were labeled as containing people or animals. This was done to avoid the known strong biases to these categories that might skew our predictions of attention control [3, 17]. (3) Only images of analog clocks were selected. This latter constraint, which was implemented by manual exclusion of images having digital clocks, was introduced because the features of analog and digital clocks are very different and this would be expected to create variability in the search behavior and reduce data quality. As a result of these exclusion criteria, the microwave-clock dataset used for training in this study consisted of 689 images containing microwaves and 1494 images containing clocks.

Because a search task requires participants to judge for each image whether the target is present or absent, the target-present (TP) images were balanced against an equal number of target-absent images (TA). TA images were selected randomly from the COCO2014 training images[2] such that: (1) none depicted an instance of the target, and (2) all depicted at least two instances of the target's siblings, where a microwave sibling was defined as an oven, a toaster, a refrigerator, or a sink object under the parent category of appliance, and a clock sibling was defined as a book, a vase, scissors, a hairdryer, a toothbrush, and a teddy bear under the parent category of indoor. This was done to discourage TA responses from being based on scene type (e.g., a city street scene would be unlikely to depict a microwave).

The large size of the dataset required that the search images for each target category be distributed over groups of searchers. For the microwave dataset, images were divided into 8 sub-groups and each sub-group was viewed by a sin-

[2]In the released dataset, there are more TA images than TP due to additional criteria being imposed on the selection of TP images for analysis.

Figure 3: Example images from the training (top row) and testing (bottom row) datasets. From left to right: a hard-to-find microwave, an easy-to-find microwave, a hard-to-find clock, an easy-to-find clock.

gle participant. Eight participants, at minimum, were therefore needed to view the entire microwave dataset. Given a final total of 27 participants in the microwave search task, each TP/TA image was searched by 3-4 different participants. For the clock dataset, images were divided into 20 sub-groups, resulting in each of the TP/TA images being searched by 1-2 different viewers based on a total of 26 participants.

Participants were verbally instructed to search for the designated target category and to make a target present or absent judgment for each image. Specifically, they pressed the "yes" button to indicate that they found the target, and the "no" button to indicate their judgment that the target does not appear in the image. Sound feedback was provided after incorrect responses. Participants viewed the images at a distance of 47cm from the monitor (resolution:1280×800), fixed by chin rest, and they were asked to fixate a central point before the display of each search image. The location of this fixation point corresponded to the center of the following search image, thereby forcing each search to begin near the image's center. The range of the search display visual angles were 12°-28.3° in width and 8°-28.3° in height. Eye position was sampled during the entire experiment at a rate of 1000 Hz using an EyeLink 1000 eye tracker (SR Research) in tower-mount configuration. Average tracker spatial error was less than 0.5° and maximum spatial error was less than 1.9°, based on calibration. Viewing was binocular, but movements of only the right eye were recorded. After removing incorrect trials and target-present trials in which the target was not fixated, 16184 search fixations remain for the images in the training dataset. Table 1 provides descriptive statistics for average number of fixations, grouped by target type and TP/TA condition. Figure 3 shows examples of easy-to-find and hard-to-find microwave and clock targets.

2.2. Testing data

A total of 40 images from COCO2014 were selected for testing, none of which overlapped with the set of training images. In addition to the criteria imposed on the selection of the training images, three more criteria were used in selecting the test set: (1) each of these images contained both a single instance of a microwave and an analog clock, (2) the size of the target was less than 10% of the image, and (3) the target could not appear at the image's center, as implemented by selecting images to avoid the center of a 5x5 grid. In our own subjective opinion, these additional criteria created what we consider to be a set of moderately-difficult images requiring active searches for the target goals, with an exemplar of each goal being in each image. Among the 40 TP test images, 27 images were from the COCO2014 training set and 13 were from the COCO2014 validation set. Similar to the TA images described for training, 40 TA images were selected from the COCO2014 validation set using the same selection criteria already described for the training dataset.

The apparatus used for behavioral data collection (EyeLink 1000), the experimental paradigm (microwave and clock categorical search tasks), and the procedural details, were all identical to what was described for the training dataset, with two exceptions: an "analog clock" was specifically mentioned in the instructions to participants tasked with searching for a clock, and there was no accuracy feedback following an incorrect response. The image stimuli were padded and resized to best fill a 1680×1050 pixel monitor (without changing image orientation), resulting in a visual angle of $54°\times35°$. A group of 30 participants searched for a microwave target and another group of 30 participants searched for a analog clock target. None of these people participated in data collection for the training dataset.

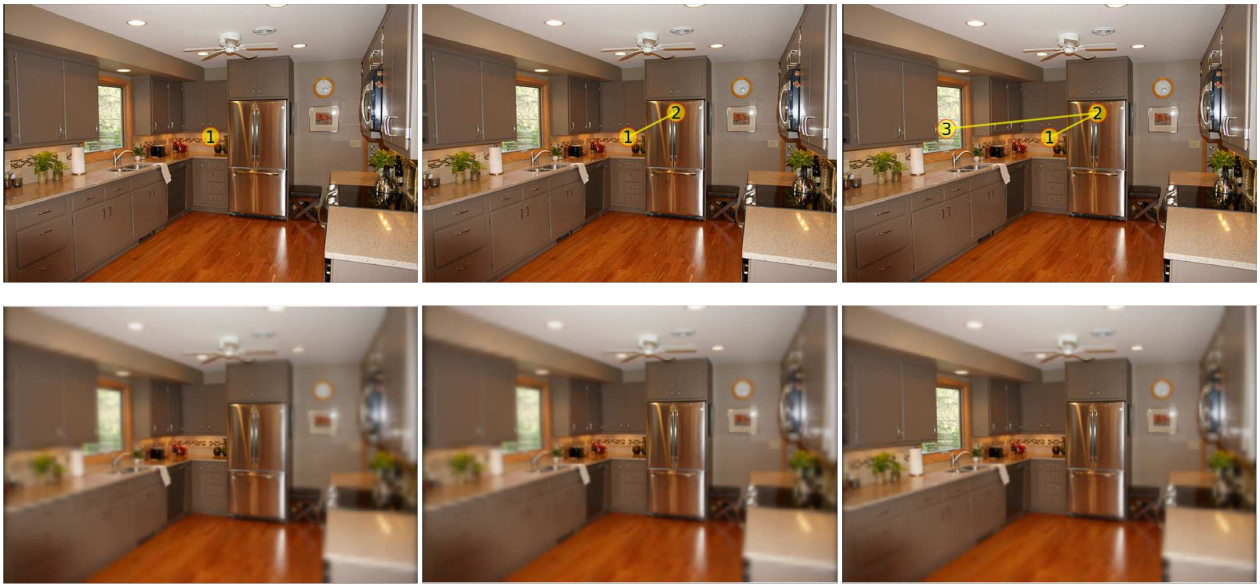

Figure 4: Example of cumulative foveated images. Top row: the locations of three fixations (columns) in a behavioral scanpath. Bottom row: the corresponding cumulative foveated images input to a model. Note that the bottom-left image shows the eccentricity-dependent blurring from a single foveated image, one based on the center fixation shown in the top-left panel. Although not intended to be an account of human information gain and retention over multiple fixations, the illustrated technique of cumulative foveated images reflects information accumulation over fixations by progressively (bottom-left to bottom-right) de-blurring a blurred foveated image based on high-resolution information obtained at each new fixation.

3. Models

Predicting the fixation scanpaths made during categorical search is extremely difficult, and in this sense our dataset is challenging because it reflects the difficulty of this task. In this context, and as a first pass, we developed several models for predicting this gaze behavior using state-of-the-art computer vision techniques. Our approach is to treat scanpath prediction as a multi-class classification problem by discretizing the image into a 10×16 grid. Each model considered here was trained as a classifier that predicts one location from among 160 possible locations, where each location corresponds to an image patch of 32×32 pixels.

The aim of these models is to predict the next fixation location conditioned on the search target and the information accumulated from previous fixations. We considered two approaches for representing this accumulated information, using either a cumulative foveated image (see Figure 4) or a recurrent neural network (RNN). The following sections detail these two approaches.

3.1. Integrating information across foveated images

Humans have a foveated retina, meaning that we have high visual acuity only at the 1° region surrounding our central vision (corresponding to our foveas), and that our vision becomes progressively more blurred with distance away from the currently fixated image location. To approximate the fact that people have high-resolution visual information centrally and lower-resolution visual information everywhere else, we obtain for each search fixation from each participant what we call a "foveated image". The bottom-left panel in Figure 4 shows one such foveated image based on the central fixation location shown in the upper-left panel. Specifically, in the current implementation each high-resolution foveal region corresponded to a 16×16 pixel image patch (about 1° of visual angle), with pixels outside this 16×16 region blurred using the method from [25].

Information maximization theories of fixation prediction [23, 26] suggest that attention control follows a greedy extraction of information from an image, which in the current context is evidence for the target goal. One way to represent this information integration in a model is to assume a build-up of high-resolution foveal images over the course of the multiple fixations made during search. We refer to this as a cumulative foveated image (Figure 4), where the high-resolution information accumulated with each fixation is a form of scanpath of reduced uncertainty (or depleted information) about whether the fixated locations contained the target. The more fixations that are made the clearer the image becomes, with "clarity" referring to a progressive reduction in the uncertainty introduced by the foveated retina [37]. This can be seen in Figure 4 by looking left to right across the bottom row; the middle and rightmost

Dataset	Category	Mean (SD)	Error (prop)
TP Training	Microwave	5.46 (2.55)	.18
	Clock	4.52 (3.50)	.15
TP Testing	Microwave	6.76 (2.12)	.09
	Clock	5.33 (1.84)	.06
TA Training	Microwave	7.95 (4.05)	.08
	Clock	11.14 (6.82)	.10
TA Testing	Microwave	14.36 (2.45)	.04
	Clock	15.85 (2.31)	.05

Table 1: The average number of fixations made before the button press for the target-present and target-absent images from both the training and testing datasets. Note that this value includes the starting fixation at the image's center. The proportion of button-press errors are also reported. Consistent with the search literature, more fixations were made in target-absent search compared to target-present. For the target-present data, microwave search took an average of one additional fixation compared to clock search regardless of dataset. Note, however, that the variability in the number of fixations was greater for the training dataset than the testing dataset. One potential reason for this is that target location and size, and the depicted number of instances of the target, were all uncontrolled variables in the training dataset (so as to have a large number of images). A consequence of this is that search task difficulty varied more for images in the training set, meaning that some training search images were very easy and others were very difficult. For the testing dataset, additional restrictions were imposed on target size and location (distance from the center) so as to sharpen the behavioral ground truth (at the expense of a far smaller number of images), with an unintended consequence of this being less variability in the number of fixations. Another potential reason for the greater variability in the training dataset is because each image in the training dataset was viewed by very few participants (1-4), whereas each testing image was viewed by 30 participants. This smaller number of viewers for each image also likely contributed to greater variability in the number of fixations.

image contains less blur than the one to its immediate left.

To model this behavior, we trained a CNN [19] to input a cumulative foveated image based on a given fixation in a scanpath and output the location of the next fixation. The architecture of our CNN was based on ResNet-50 [13] with the last average-pooling layer and fully-connected layer removed. We added two 1×1 convolutional layers and a softmax layer that maps the feature maps from ResNet-50 to a location probability map (from which we make fixation location predictions). We trained the network on the images from our training dataset, starting from a ResNet-50 [13] that was pre-trained on ImageNet [6]. We used a Winner-

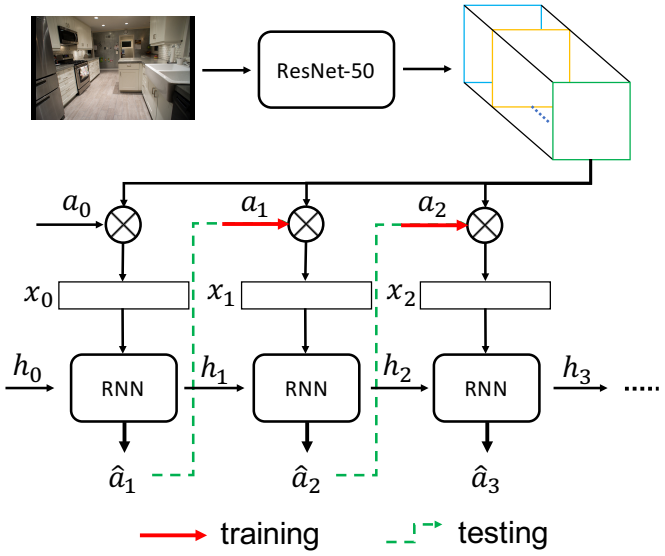

Figure 5: The general pipeline of the RNN-based methods. A pre-trained ResNet-50 is used to obtain the feature maps from an image, and at each step i we extract the feature vector $\mathbf{x}_i$ at the human ground-truth fixation location a_i. This is done both during training (shown by solid red arrows) and during testing for fixation location prediction $\hat{a}_i$ (shown by dashed green arrows). $\mathbf{x}_i$ and $\mathbf{h}_i$ are then fed into the RNN to predict the next fixation.

Take-All (WTA) strategy [4] on the location probability map, and output the location with the highest probability as the predicted fixation location.

3.2. Information integration using RNNs

An alternative approach to address the integration of information obtained over multiple search fixations is to use a Recurrent Neural Network (RNN). Figure 5 shows the pipeline of our RNN models. We use a ResNet-50, pre-trained on ImageNet, to extract feature maps from each image. At step i, a 2048-dimensional feature vector, $\mathbf{x}_i$, is extracted from the feature maps at the attended location, a_i, with $\mathbf{x}_i$ then linearly embedded to 512 dimensions and used as input to the RNN-cell at this step. The hidden state $\mathbf{h}_{i+1}$ is updated based on the current input $\mathbf{x}_i$ and the previous hidden state $\mathbf{h}_i$. The resulting new hidden state h_{i+1} is then used to predict the next action $\hat{a}_{i+1}$. The first fixation a_0 is always assumed to be at the image's center, per each participant's instruction. Treating the selection of actions as a categorical classification problem, we train the RNNs using cross-entropy loss between the behaviorally-observed fixations a_i and the predicted fixations $\hat{a}_i$. During training the human fixation a_i is used to select the feature vector $\mathbf{x}_i$, whereas at testing the predicted fixation $\hat{a}_i$ is used to generate $\mathbf{x}_i$. We consider three types of RNN models in our

Method	Shape	Direction	Length	Position
Behavioral Agreement	0.956	0.714	0.958	0.927
CNN	0.941	0.621	**0.930**	**0.895**
RNN	0.917	0.677	0.883	0.876
LSTM	0.916	**0.684**	0.879	0.876
GRU	**0.925**	0.664	0.902	0.877

Table 2: Scanpath prediction performance for microwave search.

Method	Shape	Direction	Length	Position
Behavioral Agreement	0.950	0.701	0.950	0.933
CNN	0.913	0.633	0.888	**0.893**
RNN	0.915	0.673	0.894	0.864
LSTM	0.907	**0.669**	0.879	0.863
GRU	**0.918**	0.659	**0.900**	0.883

Table 3: Scanpath prediction performance for clock search.

benchmarking, including a basic RNN [12], an LSTM [14], and a GRU [5].

3.3. Incorporating Inhibition of Return

The primate oculomotor system uses Inhibition-of-Return (IOR) to spatially tag previously attended locations for the purpose of discouraging attention from returning to regions where information has already been depleted [30]. To capture this mechanism, we incorporate IOR into all the benchmark methods by preventing gaze from attending to a visited region. We do this my masking out a 3×3 neighborhood of locations in the location probability map around each fixated location.

4. Predictions and Performance

We evaluate the predictive success of the benchmark models in two steps. First, for each model we generate a 6-fixation scanpath (using WTA and IOR) for every image in the testing dataset. The fixation length is fixed at 6 for computational convenience, but also because 6 is the approximate average length of the behavioral scanpaths in the TP testing data. Second, the search scanpaths predicted by each model are compared to the behavioral scanpaths from each participant. This scanpath comparison was done using MultiMatch [1, 7], which represents scanpath similarity in terms of five dimensions: shape, direction, length, position, and duration (see the original sources for details). However, here we exclude the duration dimension from MultiMatch because the models tested did not attempt to predict fixation duration.

Tables 2 and 3 show how well each of the benchmark models predicted the behavioral scanpaths in the microwave and clock search tasks. We also report a measure of scanpath agreement among participants (behavioral agreement), which was obtained by computing the MultiMatch scanpath similarity between each participant and all the others. This can be considered as an upper-bound for the benchmark models. As expected, scanpath similarity is highest among participants, meaning that people agree with each other more in their search behavior compared to predictions from any of the tested models. That said, benchmark model performance was overall quite good (given the performance ceiling in behavioral agreement). RNN-based models generally outperform the CNN model, suggesting that the RNN-based approaches may better capture sequential decision making in human gaze behavior.

Figure 6 shows scanpaths and action probability maps from the GRU model searching a single image for either a clock or a microwave target goal. The action probability map predicts the fixation location following the participant's next eye movement. The scanpaths are determined through peak-picking from this probability distribution, with IOR. Figure 6a shows a clock search and Figure 6b shows a microwave search, and it is this contrast that is the most interesting. While the action probability maps start the same, during the course of search they diverge to focus on image regions that are specific to the target category, namely vertical surfaces in the upper part of the image in the case of clocks and counter-top surfaces stretched horizontally across the center in the case of microwaves. So in addition to the benchmark models showing good (but not perfect) prediction of behavioral scanpaths, future work may therefore show that they are also able to capture some effects of scene context on the control of categorical search.

Finally, what is also clear from Figure 6 is that much work remains to be done on the modeling of goal-directed attention control. None of the benchmark models managed to find (fixate) the target on this test trial, regardless of whether the target was a microwave or a clock. Of course all of the models would eventually fixate the target if we have them keep making eye movements, but each of these additional eye movements lessens the model's success as a method of predicting attention control, with a very large number of fixations being tantamount to complete model failure The availability of large-scale datasets of fixations made during categorical search will finally allow deep network methods to be applied to these basic questions in goal-directed attention control.

5. Conclusions

Here we introduce the Microwave-Clock-Search (MCS) dataset, a set of images annotated with the fixations of peo-

(a) Clock (b) Microwave

Figure 6: Scanpath and action probability maps generated by the GRU model during its search for a clock (a) or microwave (b) target in the same test image. Maps are for shown for each of the fixations, ordered from left to right and top to bottom, but are limited to only the first six fixations in the scanpath. Even rows: The unfolding 6-fixation scanpath. Odd rows: Corresponding action probability maps, where a redder color indicates a higher probability of a location being fixated next.

ple searching for either microwaves or clocks. The number of these fixations depend on a number of factors (microwave or clock, training dataset or testing dataset, target present or absent), but the means range from 4.5 fixations to 15.8 fixations and we consider this to be reasonable for revealing evidence of goal-directed attention control, to the extent it exists. We used a training dataset and state-of-the-art deep network models to benchmark performance in predicting fixations in a separate testing dataset. Early qualitative and quantitative analyses suggest a promising prediction of behavioral scanpaths, but far more work is needed. Our hope is that these models will form useful baselines against which novel methods of predicting goal-directed attention control can be compared.

Acknowledgements. This work was supported by NSF Awards IIS-1763981 and CNS-1718014, the Partner University Fund, the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe.

References

[1] Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392, 2015.

[2] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015.

[3] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, pages 241–248, 2008.

[4] Zhenzhong Chen and Wanjie Sun. Scanpath prediction for visual attention using ior-roi lstm. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 642–648. AAAI Press, 2018.

[5] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[7] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012.

[8] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009.

[9] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. www.pascal-network.org/challenges/VOC/voc2012/workshop/, 2012.

[10] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1105–1108. ACM, 2014.

[11] Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.

[12] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[16] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.

[17] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[18] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] Y. LeCun, B. Boser, J. S. Denker, and D. Henderson. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[21] Justin T Maxfield, Westri D Stalder, and Gregory J Zelinsky. Effects of target typicality on categorical search. *Journal of vision*, 14(12):1–1, 2014.

[22] Justin T Maxfield and Gregory J Zelinsky. Searching through the hierarchy: How level of target categorization affects visual search. *Visual Cognition*, 20(10):1153–1163, 2012.

[23] Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387, 2005.

[24] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014.

[25] Jeffrey S Perry and Wilson S Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *Human vision and electronic imaging VII*, volume 4662, pages 57–70. International Society for Optics and Photonics, 2002.

[26] LW Renninger, JM Coughlan, P Verghese, and J Malik. An information maximization model of eye movements. In *Advances in Neural Information Processing Systems*, 2005.

[27] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition? In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[28] Joseph Schmidt and Gregory J Zelinsky. Search guidance is proportional to the categorical specificity of a target cue. *The Quarterly Journal of Experimental Psychology*, 62(10):1904–1914, 2009.

[29] BW Tatler, MM Hayhoe, MF Land, and DH Ballard. Eye guidance in natural vision: Reinterpreting. 2011.

[30] Zhiguo Wang and Raymond M Klein. Searching for inhibition of return in visual search: A review. *Vision research*, 50(2):220–228, 2010.

[31] Zijun Wei, Hossein Adeli, Minh Hoai, Gregory Zelinsky, and Dimitris Samaras. Learned region sparsity and diversity also predict visual attention. In *Advances in Neural Information Processing Systems*, 2016.

[32] Zijun Wei and Minh Hoai. Region ranking SVMs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[33] Niklas Wilming, Selim Onat, José P Ossandón, Alper Acik, Tim C Kietzmann, Kai Kaspar, Ricardo R Gameiro, Alexandra Vormberg, and Peter König. An extensive dataset of eye movements during viewing of complex images. *Scientific data*, 4:160126, 2017.

[34] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[35] Hyejin Yang and Gregory J Zelinsky. Visual search is guided to categorically-defined targets. *Vision research*, 49(16):2095–2103, 2009.

[36] Chen-Ping Yu, Justin T Maxfield, and Gregory J Zelinsky. Searching for category-consistent features: A computational approach to understanding visual category representation. *Psychological science*, 27(6):870–884, 2016.

[37] Gregory J Zelinsky. A theory of eye movements during target acquisition. *Psychological review*, 115(4):787, 2008.

[38] Gregory J Zelinsky and James W Bisley. The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, 1339(1):154–164, 2015.

[39] Gregory J Zelinsky, Yifan Peng, Alexander C Berg, and Dimitris Samaras. Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision*, 13(3):30–30, 2013.

[40] Gregory J Zelinsky, Yifan Peng, and Dimitris Samaras. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of vision*, 13(14):10–10, 2013.