Neurocomputing 149 (2015) 788–799

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

What do eyes reveal about the mind? Algorithmic inference of search targets from fixations

Ali Borji [a,*], Andreas Lennartz [b], Marc Pomplun [b,1]

[a] Department of Computer Science, University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, CA 90089, United States
[b] Department of Computer Science, University of Massachusetts at Boston, 100 Morrissey Boulevard, Boston, MA 02125-3393, United States

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form
23 July 2014
Accepted 24 July 2014
Communicated by Hantao Liu
Available online 21 August 2014

Keywords:
Visual attention
Eye movements
Visual search
Intent decoding
Mind reading
Yarbus

ABSTRACT

We address the question of inferring the search target from fixation behavior in visual search. Such inference is possible since during search, our attention and gaze are guided toward visual features similar to those in the search target. We strive to answer two fundamental questions: what are the most powerful algorithmic principles for this task, and how does their performance depend on the amount of available eye movement data and the complexity of the target objects? In the first two experiments, we choose a random-dot search paradigm to eliminate contextual influences on search. We present an algorithm that correctly infers the target pattern up to 50 times as often as a previously employed method and promises sufficient power and robustness for interface control. Moreover, the current data suggest a principal limitation of target inference that is crucial for interface design: if the target pattern exceeds a certain spatial complexity level, only a subpattern tends to guide the observers' eye movements, which drastically impairs target inference. In the third experiment, we show that it is possible to predict search targets in natural scenes using pattern classifiers and classic computer vision features significantly above chance. The availability of compelling inferential algorithms could initiate a new generation of smart, gaze-controlled interfaces and wearable visual technologies that deduce from their users' eye movements the visual information for which they are looking. In a broader perspective, our study shows directions for efficient intent decoding from eye movements.

1. Introduction

Eye movements can reveal a wealth of information about the complex cognitive states of the mind. They carry information that is diagnostic of the task an observer is trying to perform [16,85,20,24,35,2,38,6]. Yarbus, in his seminal work in 1967, reported that observers' fixation patterns during free viewing of a painting were dramatically different than when different questions were given [85]. While the allocation of attention is often task-driven, it can also be guided by bottom-up and stimulus-driven cues [80,42,40,64,39,4,3,10]. Normal vision employs both processes simultaneously to control overt and covert shifts of attention.

There is a rich collection of literature that discusses the role of oculomotor behavior in tasks as diverse as reading [69,22], pattern copying [1], portrait painting [55], visual search [80,84,88], tea making [44], sandwich making [32], fencing [29], cricket [46], squash [21], billiards [23], juggling [43], activity recognition [15,50,65,25], and game playing [9,7,11]. See [47] for a review of eye movements in natural vision tasks. Some general underlying principles of gaze guidance have been discovered. For example, it is known that eye movements follow the road tangent in driving [45], some saccades occur to avoid obstacles (predictive saccades in walking [54]), and eye movements are sensitive to the value of visual items [59]. Eye movements are also indicators of abstract thought processes, for instance in arithmetic and geometric problem solving [18], list sorting, and mental imagery [53]. These findings highlight the intricate links between the mind, the body's actions, and the world around us. This active aspect of vision and attention has been extensively investigated in the context of natural behavior. Please see [1,33,78,74,57,48,47,39,5] for reviews.

Some computational models have been proposed to quantify gaze behavior, though their generalizations across tasks remain limited. Examples of top-down models of gaze control include HMM models of fixation prediction in reading (E–Z reader model by Reichle et al. [70], Mr. Chips model by Legge et al. [49]), a model of minimizing local uncertainty in object classification [71], a reward maximization framework to coordinate basic visio-motor routines to perform a complex task using reinforcement learning [77], Bayesian models of gaze control (e.g., [86,72,8]), and pattern classification models [9,7]. In addition, a myriad of

[*] Corresponding author. Tel.: +1 310 993 3988.
E-mail addresses: borji@usc.edu (A. Borji), andreas.lennartz@gmail.com (A. Lennartz), marc@cs.umb.edu (M. Pomplun).
[1] Tel.: +1 617 287 6443.

http://dx.doi.org/10.1016/j.neucom.2014.07.055
0925-2312/© 2014 Elsevier B.V. All rights reserved.

bottom-up models exist for predicting where observers look when engaged in free-viewing of pictures of natural scenes (see the review by Borji and Itti [5]).

Despite the enormous amount of past research on understanding the mechanisms of gaze control, less systematic effort has been made so far to predict intents from fixations. The majority of studies have qualitatively analyzed the difference between eye movement patterns of observers viewing natural scenes under different questions (e.g., [24,6]). Some researchers, conducting quantitative analyses, have reported that it is possible to decode the task from eye movements while some others have argued against it. For example, Henderson et al. [35] recorded eye movements of 12 participants while they were engaged in four tasks over 196 scenes and 140 texts: scene search, scene memorization, reading, and pseudo reading. They showed that the viewing tasks were highly distinguishable based on eye movement features in a four-way classification (decoding accuracy above 80%). In contrary, Greene et al. [28] did an experiment in which they recorded eye movements of observers when viewing scenes under four questions: memorize the picture, determine the decade in which the picture was taken, determine how well the people in the picture know each other, and determine the wealth of the people in the picture. They were able to decode image and observer's identity from eye movements above chance level, but failed to predict the viewer's task (see Fig. 4 in Greene et al.'s paper). Borji and Itti [6] were later able to decode observers' task on this data as well as on the original question of Yarbus. Several successful attempts have been made in the past to learn about human cognition such as predicting search targets [68,30], decoding stimulus category [31,60,12], predicting relative magnitude of a randomly chosen number by a person [51], predicting events [65,15], predicting an observer's category of clinical condition [81], and task decoding [85].

The current study addresses the challenging problem of intent decoding – predicting what target an observer is looking for from his eye movements. Some scientific findings show promising directions in this regard. For example, it is known that during visual search, our attention and eye movements are biased by visual information resembling the target (e.g., [56,66,84,68]), causing the image statistics near our fixated positions to be systematically influenced by basic visual features of the target ([68,66]). One study also found that the type of object sought, of two possible categories, can be inferred from search statistics [87]. However, the existing approaches have not considered strategies beyond using elementary search statistics [68]. Furthermore, current methods have not been tested for target decoding on natural scenes.

Our work focuses on designing powerful search target inference algorithms from eye movements recorded during visual search. Visual search is an important task as it is one of the main ingredients of complex daily life tasks. Two important application domains of such target prediction algorithms are interface design (e.g., smart webpages) and wearable visual technologies. If target inference becomes possible for a large set of candidate objects, a new generation of smart, gaze-controlled human–computer interfaces could become reality [36,75]. Gaining information about an interface user's object of interest, even in its absence, would be invaluable for the interface to provide the most relevant feedback to its user. In a broader perspective, our study shows directions for efficient intent decoding from eye movements.

2. Visual search experiments

We conduct three experiments to explore the potential of algorithmically inferring the search target from a searcher's visited patterns. In the first two experiments, we choose a random-dot search paradigm to eliminate contextual influences on visual search (see Fig. 1 for example scenes). The proposed techniques could also be applied to the local feature vectors of any type of display.

Search in natural scenes is different from looking for targets in random-dot patterns since several other factors, in addition to target features, are involved. Those factors include global scene context [79], background clutter [73], object-semantic dependencies [37], and spatial priors [63]. In the third experiment, to investigate informativeness of fixated image patches, we attempt to predict the search target in natural scenes only from image patches centered at fixations.

Before proceeding to algorithms, we define two terms: *fixated patterns* is the set of all patterns that a subject visits while viewing the search array, and *generated patterns* is the set of patterns that we generate from fixated patterns by considering windows around them (by sliding a 3×3 window around each fixated pattern).

2.1. Experiments 1 and 2: searching for a target on a synthetic background

In these experiments, subjects searched a large random-dot array for a specific 3×3 pattern of squares in two (Experiment 1) or three (Experiment 2) luminance levels while their eye movements were measured. Our aim was to devise algorithms that received a subject's gaze-fixation positions and the underlying display data and inferred the actual target pattern with the highest possible probability. Fixation and display data from the actual target pattern in the search display were excluded, because the disproportionate fixation density at the end of a search would have made target inference trivial. A variety of inferential algorithms and classifiers were devised and tuned based on ten subjects' gaze-position data and evaluated on another ten subjects' data for each experiment. The current paradigm was well-suited

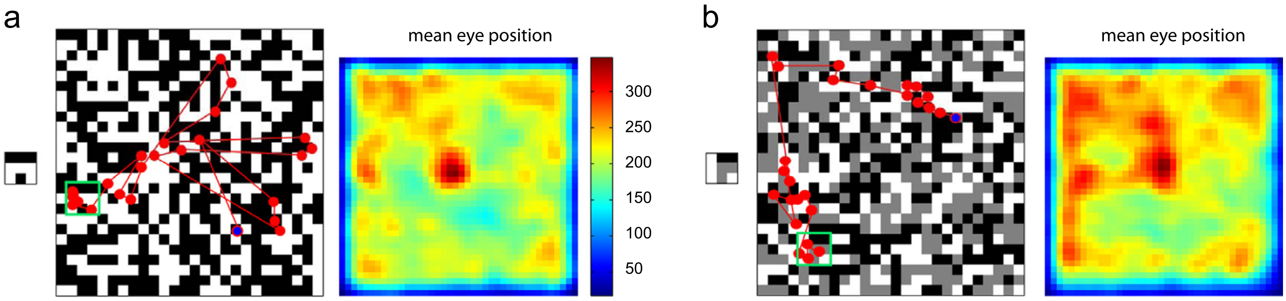

Fig. 1. Search targets and cut-outs from the corresponding visual search displays in (a) Experiment 1 and (b) Experiment 2 with human subjects' scanpaths superimposed on them. Actual displays consisted of 40×40 squares. Red discs indicate fixation positions, consecutive fixations are connected by straight lines, and the initial fixation is marked with a blue dot. A green square indicates the position of the target in the search display. Mean eye position over all trials for each experiment is also shown. *Fixated patterns* is the set of all patterns that a subject visits while viewing the search array, and *generated patterns* is the set of patterns that we generate from fixated patterns by considering windows around them. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

for a first quantitative exploration of this field, because it minimized the influence of semantic factors on eye movements (cf. [34]) and supplied fixed numbers of equally probable target patterns, $2^9 = 512$ in Experiment 1 and $3^9 = 19,683$ in Experiment 2. At the same time, this paradigm challenged the algorithms to the extreme, not only due to these huge numbers of target candidates, but also because they were not shown as discrete objects but formed a contiguous pattern whose elements barely exceeded the spatial resolution of the eye tracking system.

The same twenty subjects aged 19–36 with normal or corrected-to-normal vision participated in each experiment after giving informed consent. Their eye movements were measured using a head mounted eye tracker (EyeLink-II, SR Research, Mississauga, Canada) with an average accuracy of 0.5° and a sampling rate of 500 Hz. At the start of each trial in Experiment 1, subjects were presented with their search target – a 3×3 array of squares (width 0.6° of visual angle), each of which was randomly chosen to be either black ($1.2\,\text{cd/m}^2$) or white ($71.2\,\text{cd/m}^2$). In Experiment 2, a third luminance level (gray, $36.2\,\text{cd/m}^2$) was added. Subjects had 6 s to memorize this pattern before it was replaced with the search display consisting of 40×40 squares of the same size and luminance levels as those in the target. Each search display contained the target pattern exactly once (Fig. 1). Subjects were instructed to find the target as quickly as possible, then fixate on it and press a designated button to terminate the trial. If the distance between gaze position and target object during the button press was less than 1°, successful target detection was counted. If no response occurred within 60 s after the onset of the search display, the trial also terminated. In each experiment, every subject performed a total of 30 trials, and during all of these trials the search target remained the same. We analyzed the data of all trials regardless of success or failure of a trial. The reason for including trials without target detection is that we were not interested in the observers' ability to detect the target but in their efforts of finding it. All subjects found their target in some of the trials, and there was no indication that they were not performing their assigned task at any point during the experiment.

Fig. 1 shows a bias in mean eye position towards the upper-left quadrant of the display. This bias seems to be present in both experiments but more prevalent in Experiment 2. It is likely a consequence of systematic scanning of the display in reading direction, which was left-to-right and top-to-bottom for all subjects. Since trials often timed out before the display had been completely searched, more fixations were located in the upper-left quadrant of the display than in the other quadrants. The reading direction effect is typically more pronounced in more difficult search tasks [67], and therefore the eye-movement bias is more pronounced in Experiment 2.

2.2. *Experiment 3: searching for an object in a natural scene*

We used the data collected by [83], consisting of 11 full-color, 3D rendered images of real-world scenes and containing 15 singleton targets, i.e., only one object resembling the target was present in each scene (see Fig. 8a). Six subjects with normal or corrected-to-normal vision participated in this experiment (mean age=25; SD=5). Scenes were displayed on a 19-in computer screen (resolution 1024×768, 100 Hz) subtending visual angles of 37° (horizontal) and 30° (vertical) at a viewing distance of 65 cm. Eye movements were recorded with an EyeLink-1000 desktop mount system (SR Research, Canada) at a sampling rate of 1000 Hz. Prior to each search trial, the target object was specified by a word presented in the center of the scene. Participants were instructed to search for the object as fast as possible and, once found, to press a button of a joystick while fixating the object. Participants were asked to search, one after another, for 15 different objects in the same scene (i.e., 15 search trials). The search scene was then replaced by another scene for the next 15 trials, and so on, for a total of 165 trials.

3. Algorithms for inferring search targets

Our development and evaluation of several inferential algorithms on synthetic patterns resulted in the discovery of two particularly powerful mechanisms, whose combination outperformed all other methods (including a baseline method from [68]) over the first two experiments without modifying their parameters between experiments. We also checked the generality of our results using pattern classifiers (Support Vector Machines and Naive Bayes) over Experiment 3 and sought whether a top-down biasing phenomenon exists in searching for a target on complex natural scenes with several objects and background clutter. Thus, we examined a total of five search target prediction methods. In Experiments 1 and 2, algorithm parameters were first tuned from the data of ten subjects (train set) and were then evaluated on the other ten subjects' data (test set). All results reported in this paper pertain to the test set. In Experiment 3, evaluation was performed using cross-validation by splitting data into train and test sets.

We first compiled all fixated patterns from all search trials of each subject into a large dataset (data of each subject taken individually; one dataset per subject). For each fixated square, a 3×3 window was placed over it nine times so that each of its squares landed on the fixated square once. This technique was chosen to account for the uncertainty of subjects in landing their eye movements and/or eye tracker error. All the generated patterns in this way were added to the dataset of each subject. Fig. 2 shows the number of generated patterns for each subject taken individually. Overall there were 293,105 generated patterns for all ten test subjects, taken together, in Experiment 1 and 313,174 in Experiment 2.

To evaluate the performance of algorithms, we repeatedly sampled (without replacement) from fixated (generated) patterns of each subject and formed a pool. We then progressively ran over all patterns in this pool, calculated and updated the statistics from this set, and used these statistics to estimate the target of each fixated pattern. The maximum number of fixations (pool size), allowing tractable processing, for this analysis was chosen to be 1500 in Experiment 1 and 800 in Experiment 2. We ran each algorithm with randomly chosen pools many times to obtain a reliable estimation of accuracy. Algorithms were run on each

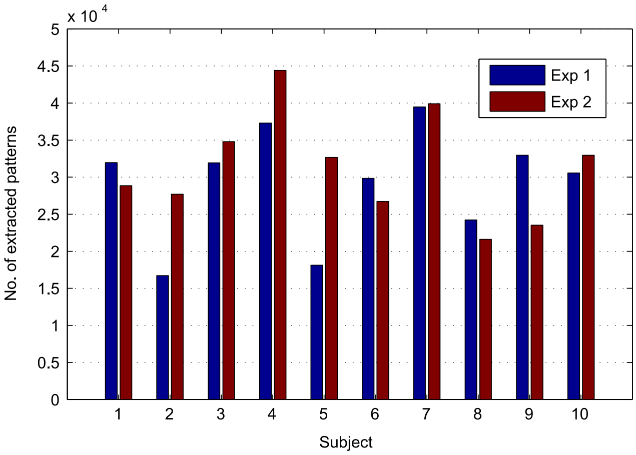

Fig. 2. Extracted patterns over ten subjects in Experiments 1 and 2. Each fixation, off the image border, generates 9 patterns.

subject separately (since they searched for different targets) and then accuracies were averaged. The pseudo code of this process is illustrated in Algorithm 1.

Algorithm 1. Search target prediction algorithm [num_subjects=10; num_runs=10,000 and 15,000 in Exp. 1 and Exp. 2, respectively; num_fixations is the variable].

```
Input: Generated patterns for each subject/search target
    pattern
Output: Mean target prediction accuracy
for s = 1⋯num_subjects do
    for r = 1⋯num_runs do
        Pool = Randomly choose num_fixations patterns from
        generated patterns of subject s
        for f = 1⋯num_fixations do
            Compute and update average luminance of each
            fixated 3 × 3 pattern in Exp. 1 (over the entire data; not
            pool) or number of votes in Exps. 2 and 3 in pool
            Predict the search target
        end for
        Compute mean accuracy over all fixations
    end for
    Compute mean accuracy over all runs
end for
Compute and report mean accuracy over all
subjects
```

3.1. Gaze-centered feature map algorithm

The first algorithm, here termed as *gaze-centered feature map*, is adapted from a previous study by Rajashekar et al. [68]. In that study, the statistical distribution of display luminance in a window centered on a subject's fixation positions was measured and in some cases found to roughly resemble the search target. To apply this method to the current task, we computed the frequency of each feature (e.g., black, gray, and white in Exp. 2) in each square (across all 3 × 3-square fixated windows over the entire data) and subtracted the average frequency of that feature across the nine squares as we increased the number of fixations. The feature with the highest value in each square entered the estimated target pattern.

3.2. Pattern voting algorithm

Our first newly developed technique, *pattern voting*, is based on the assumption, derived from an earlier study by Shen and Reingold [76], that the strongest attractors of observers' eye movements during search are local patterns that are very similar to the search target. We operationally defined the similarity between two 3 × 3 patterns as the number of matching features in corresponding squares, resulting in a range of similarity values from zero to nine (i.e., Hamming distance). The voting algorithm keeps score of the votes for every possible 3 × 3 pattern. Each time, the patterns whose similarity to the fixated pattern in the window is eight (high-similarity patterns) receive one vote. Identical patterns (similarity nine) do not receive votes for the benefit of a 'fair' evaluation, since neither the actual target nor the fixations on it are visible to the algorithm. The pattern receiving the most votes is the estimated target pattern. In case of ties in votes, we randomly selected one of the patterns with highest votes (see Appendix A).

3.3. Weighted pattern voting algorithm

Interestingly, placing only the window center over fixated squares or weighting this center position more heavily leads to reduced performance of the voting algorithm. While this effect may partially be due to noise in gaze-position measurement, it is also possible that subjects do not always fixate on the center of a suspected target. Depending on how they memorize the target, their gaze may be attracted by a specific position within similar patterns – a 'gaze anchor' position from where they compare the local pattern with the memorized one. If we could estimate the most likely gaze anchor positions, we could improve the pattern voting algorithm by assigning greater weights to the votes received at the corresponding window positions relative to fixation. These window positions should be indicated by greater consistency of their high-similarity patterns, that is, stronger preference of some patterns over others. Preliminary experimentation showed that effective weighting can be achieved by computing separately for the nine window positions the votes for individual patterns as above, divide them by the average number of votes for that position, and apply an exponent. The final score for a pattern is the sum of its weights across the nine positions, and the highest score determines the estimated target pattern. The exponent, which rewards high frequencies of patterns in specific positions, should increase when more gaze samples are provided in order to exploit the greater signal-to-noise ratio. The final *weighted pattern voting* algorithm computes the final score s_n for pattern n as follows:

$$s_n = \sum_{r=1}^{R} \left(\frac{N.v_{r,n}}{V_r}\right)^z, \quad z = \ln\left(e + \frac{V_r}{c}\right), \texttt{for } n = 1, \cdots, N \tag{1}$$

where N is the total number of patterns (512 or 19,683 in this study), R is the number of distinct window positions relative to fixation (here, $R=9$), $v_{r,n}$ is the number of votes given by the pattern voting algorithm to pattern n in window position r, V_r is the sum of votes for all N patterns in position r (i.e., $V_r = \sum_{n=1}^{N} v_{r,n}$), and c is a constant whose optimal value was found near 600 for both experiments, based on the preliminary data.

3.4. Pattern classifiers

The number of possible target classes in the synthetic patterns in the first two experiments is relatively large (512 in Experiment 1 and 19,683 in Experiment 2). For the majority of these classes (e.g., 502 classes in Experiment 1), we do not have labeled data. This makes training traditional machine learning classifiers infeasible on this data. Here, to explore the specificity of searched patterns for each target and check whether search for a target indeed biases the viewing behavior, we utilize two pattern classifiers: Support Vector Machines (SVMs, see [17]) and Naive Bayes. Each classifier is trained on a subset of data and tested on the remaining part of the data (i.e., cross validation). We report accuracies of these classifiers over all three experiments.

4. Target decoding results

4.1. Experiments 1 and 2: target inference on synthetic patterns

We first analyze the degree to which eye movements convey information regarding the search target compared to randomly fixated locations. Let $s(t_i, T) = 1 - h(t_i, T)$ represent the similarity between the fixated pattern t_i and search target T, where $h(,)$ is the Hamming distance. In Fig. 3a, we show the similarity between the most frequent fixated patterns (excluding first and last fixations) and the target. Taking the cumulative mean of the similarity

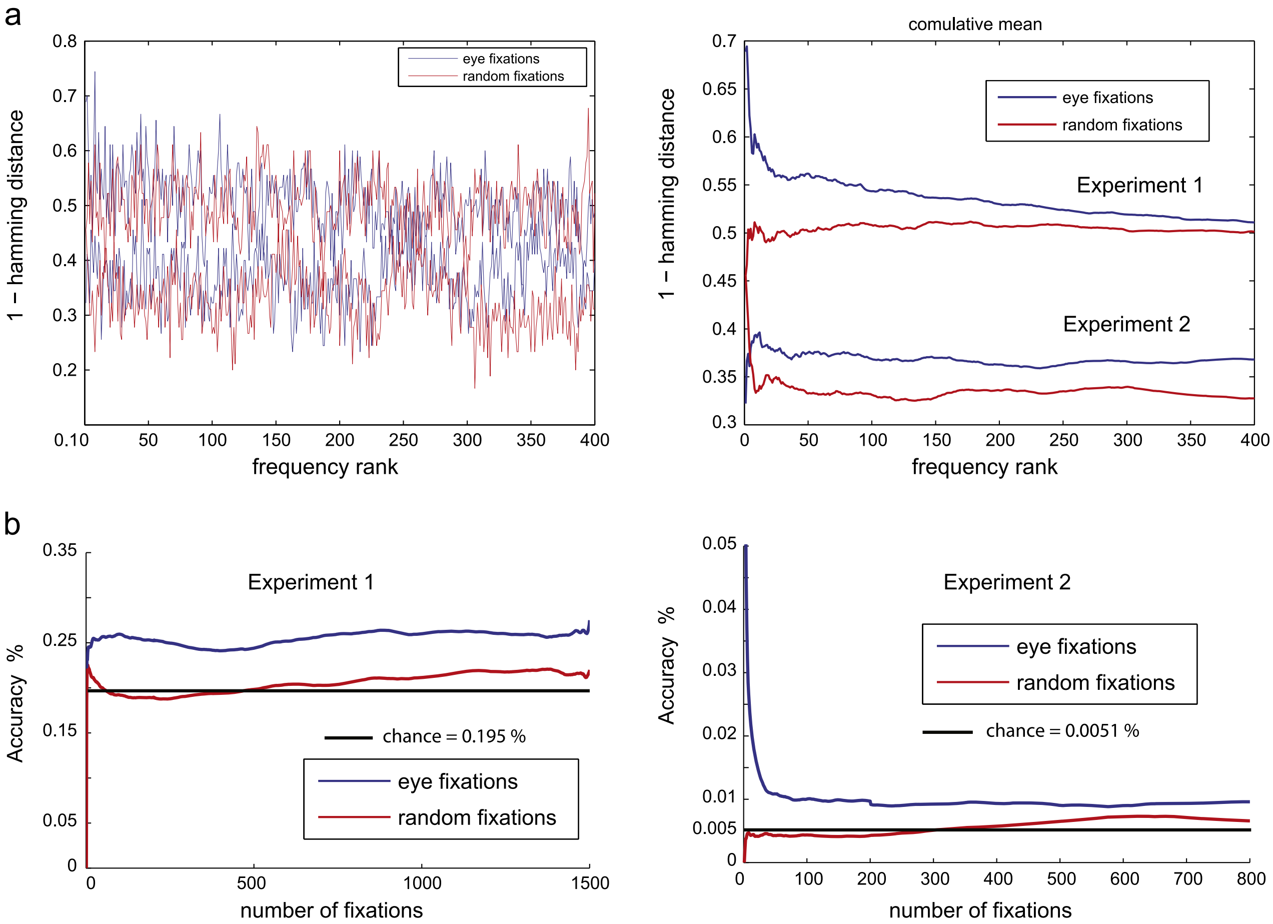

Fig. 3. (a) Similarity (1 − Hamming distance) between the 400 most frequently fixated patterns and the target. Occurrences of fixated patterns were first counted and then sorted by frequency. Results are calculated for all trials of each subject and then are averaged across all subjects. Right panel shows the running average. (b) Decoding accuracy of the gaze-centered feature map for Experiments 1 and 2.

measure (Fig. 3a, right) shows that highly fixated patterns resemble the target more strongly than less frequent ones. In other words, subjects tend to look more often at patterns that are similar to the target. Note that s values are higher in Experiment 1 than in Experiment 2 due to lower patten variability.

Decoding accuracy of the *gaze-centered feature map* is shown in Fig. 3b for both experiments. Higher prediction accuracy was obtained with human fixations than with random fixations, which were used to further control for systematic biases in the stimuli. Note that the actual chance level is very low (0.195% and 0.0051%, respectively) so doing anything better than chance with the small patches (about 0.6° visual degrees) would indicate a significant success. The gaze-centered feature map led to above chance target prediction accuracies of 0.27% and 0.01% for Experiments 1 and 2, respectively.

The chart in Fig. 4 illustrates that *pattern voting* clearly outperforms the gaze-centered feature map in Fig. 3b. In Experiment 1, even after only 20 fixations (randomly sampled from the pool; about 5 s of search in a search trial), the voting algorithm's probability of picking the correct target is already 4.8 times above chance level, while it is only 1.2 times above chance for the feature map. After approximately 300 fixations, the *weighted pattern voting* starts surpassing the basic voting algorithm and maintains a steeper increase until the final 1500 fixations, where its performance reaches 13.3%, outperforming the voting algorithm (10.4%, $p < 0.01$), which in turn exceeded the performance of the gaze-centered feature map ($p < 0.001$; Fig. 4a). A similar pattern of results was found for Experiment 2, with maximal accuracy of 0.052% for voting and 0.065% for weighted voting based on 800 fixations (Fig. 4b). In both experiments, pattern voting algorithms perform significantly better than the gaze feature map approach (ps < 0.05). See also Appendix B for results using other variants of voting methods.

Even if we compensate for the difference in pattern set size (by dividing performance by chance level), weighted pattern voting still performs clearly better in Experiment 1 than in Experiment 2, as indicated by greater performance-to-chance level proportion (68.2 in Exp. 1 vs. 13 in Exp. 2 at 800 fixations; 34.07 in Exp. 1 vs. 6.5 in Exp. 2 with regard to random fixation chance-level at 800 fixations), and sensitivity d' (2.54 vs. 0.92, respectively) according to signal detection theory, $p < 0.01$, for 1800 fixations.[2] If the reason for this discrepancy was poorer memorization of the more complex target patterns in Experiment 2 and, as a result, greater noise in the guidance of eye movements, then subjects should detect the target less often than they do in Experiment 1. However, the mean target detection rate (i.e., subject success rate) is 43% in Experiment 1 and 47.3% in Experiment 2. Another possible explanation is that the higher target complexity leads to subjects' eye movements being guided by only a part of the target pattern,

[2] To compare the inferential performance of algorithms between decision spaces of different sizes, we employed the sensitivity measure d' for the situation in which a technical device or human observer has to make a choice among a known number of alternatives [52]. Although this measure assumes independence of signals, which is not warranted in the present scenario, it provides a useful approximation that has been applied to similar problems before [52]. In the subpattern analysis (Fig. 5a), we further make the simplifying assumption that all subpatterns of a target are fixated with the same probability.

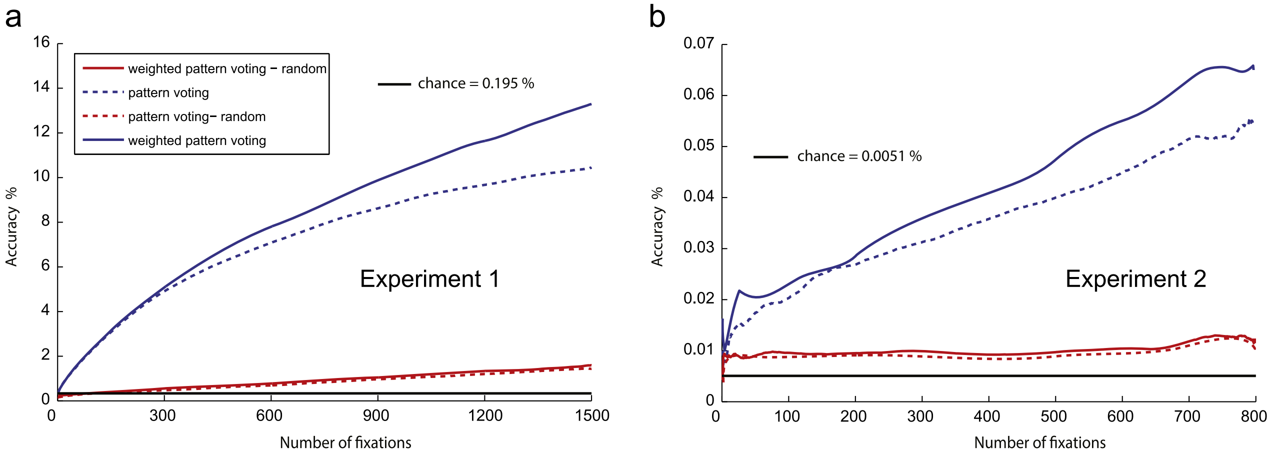

Fig. 4. Comparison of inferential performance of pattern voting algorithms proposed here. Performance is measured as the probability of correctly inferred target objects as a function of the number of gaze fixations provided to the algorithms in (a) Experiment 1 and (b) Experiment 2. This probability was approximated by repeated resampling (10,000 and 15,000 times for Experiments 1 and 2, respectively) of subjects' fixation data. Notice that the number of potential target patterns is 512 in Experiment 1 and 19,683 in Experiment 2.

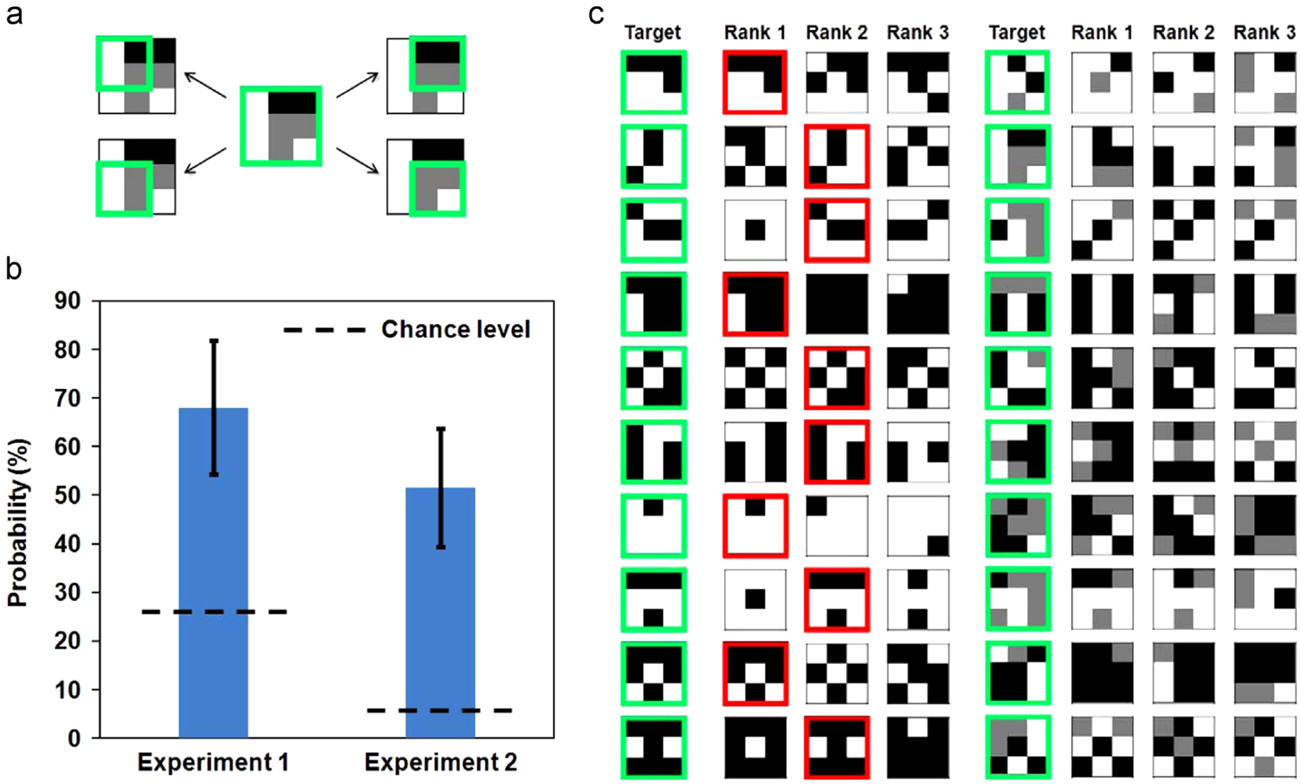

Fig. 5. Analysis of target subpattern frequencies. (a) Each target is decomposed into four 2 × 2 subpatterns. (b) Probability of any of the four target subpatterns to receive the most fixations among all 2 × 2 subpatterns (16 patterns in Experiment 1 and 81 patterns in Experiment 2). Error bars indicate standard error of the mean across ten subjects. (c) Actual targets (green frame) and the 3 patterns ranked highest by the weighted voting algorithm (left: Exp. 1, right: Exp. 2). Actual target appearing in the first 3 ranks are marked by red frames. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

and whenever this part is detected, a complete verification of the local pattern is conducted. To test this hypothesis, we resampled 1800 fixations (since the minimal dataset across subjects and experiments included 1821 fixations) to rank all 2 × 2 patterns according to their frequency of being fixated, and calculated the probability that any of the four 2 × 2 subpatterns of the target (Fig. 5b) was the top-ranked one. While the absolute hit rate (i.e., algorithm performance) does not differ statistically between Experiments 1 and 2 (68.1% vs. 51.6%, respectively), $p > 0.3$, both the hit rate-to-chance level proportion (2.72 vs. 10.44, respectively) and sensitivity d' (0.65 vs. 1.19, respectively), are greater in Experiment 2, $p < 0.01$, supporting our hypothesis (Fig. 5b).

Fig. 5c illustrates the top ranking choices made by the weighted pattern voting algorithm. Actual target objects appearing in the first three ranks are marked by red frames. While all target patterns in Experiment 1 occupy either rank one or two (out of

512 candidates), the average rank of the target patterns in Experiment 2 is 1514 (out of 19,683 candidates).

Table 1 summarizes search target inference results over the first two experiments (Fig. 6).

We now report accuracies of *Naive Bayes* (*NB*) and *multi-class SVM* for search target/subject inference with a linear kernel using generated patterns (linearized 3 × 3 patterns). Note that our goal here is not target inference (out of all possible search targets) but rather subject identity inference. We need this alternate comparison where the goal is to predict one of ten subjects for whom we have training data. Recall that we do not have enough labeled data for all possible search targets to run traditional machine learning classifiers. Chance level here is 10%, which is much higher than in the original experiments.

We have data for only 10 classes (10 subjects, each one looking for a different target). Here, we drew $n \in \{2, 5, 10, 100, 200, 500, 1000, 5000\}$ patterns uniformly randomly from each subject's generated patterns and pooled them across all 10 subjects to build a train set. Each pattern was accompanied by its target label which was the same as the subject's identity, since each subject searched for a unique target. To generate the test set, we followed the same procedure, except now each time we picked n patterns from the rest of the patterns (i.e., excluding training patterns). The above process was repeated 200 times and results were averaged. Note that chance level was at 10% (1 out of 10 targets). Fig. 5 shows decoding results for human generated and randomly generated patterns. Including more fixations increased accuracy up to 16.58% for Exp. 1 (14.46% using Naive Bayes) and 15.8% for Exp. 2 (13.3% using Naive Bayes), both significantly above chance (t-test, $p < 0.005$ using both classifiers starting from 10 fixations). Decoding performance was at chance level with random fixations. This result indicates that fixated locations were more similar to the target pattern, conceivably due to a top-down attention biasing mechanism [56,66,84,68,58]. Note that this outcome was not caused by the difference in statistics of search arrays (i.e., different random dot patterns over subjects), as classification accuracy with random patterns was not significantly better than chance.

Table 1
Summary results of Experiments 1 and 2 for search target inference. Accuracies are in percentages. Stars indicate significance vs. voting algorithm ($p < 0.01$ in both experiments). In both experiments, both voting algorithms perform significantly better than the gaze feature map approach (ps < 0.05).

Parameter/Accuracy	Exp. 1	Exp. 2
No. of classes	$2^9 = 512$	$3^9 = 19{,}683$
No. of fixations in pool	1500	800
Gaze-centered map	0.27	0.01
Voting	10.4	0.052
Weighted voting	13.3*	0.065*
Chance	0.195	0.0051

Strengthen this observation by clarifying for the reader that the test set was derived from a different set of target patterns than the training set. Therefore, it makes sense that the subject classifier will fail when the gaze patterns are influenced more by target identity than by subject identity.

4.2. Experiment 3: target inference on natural scenes

We investigated the influence of two parameters in target decoding accuracy on data from Section 2.2 (see Fig. 7) including (1) number of saccades from 1 to 15, and (2) attention window size (patch size n) from 5 × 5 to 513 × 513 (i.e., $2^n + 1$ for $n = 2 \ldots 9$). First and last saccades as well as all saccades inside the target object region (object boundaries were annotated) were discarded. A classifier was trained on each image separately and then its average performance over all 11 images was measured. Following a leave-one-out procedure, we first trained a *multi-class SVM* with RBF kernel from data of 5 subjects to map randomly selected observer-fixated patches to the target object. The resultant classifier was then applied to another set of randomly selected fixated patches from the remaining subject. We exploited three types of features to encode the image content at fixated locations: (1) the concatenated histograms of red, green, and blue pixels in RGB color space, (2) gist features by Oliva and Torralba [62] which have proven to be highly relevant to scene and object recognition, and (3) local binary patterns (LBP) by Ojala et al. [61].

Fig. 8b shows that target decoding accuracy rises with increasing number of fixations. Further, increasing the window size rises the decoding performance up to size 257 (129 for RGB Hist) and then drops. This might be because the average size of annotated objects was approximately 256 × 256 pixels, or because the size of the subjects' attentional focus was in the same range. As a control, we also used raw fixation locations (i.e., augmented (x, y) coordinates) as features for target prediction (solid black curve in Fig. 8b). Surprisingly, this simple feature results in significantly above-chance accuracy of 22.67% ($p < 0.05$) indicating systematic differences in eye movements in searching for different objects. It should be noted, however, that fixation location implicitly encodes local image features and also semantic factors; for instance, when searching for a bicycle, observers would direct their attention to the depicted street level rather than the roofs of houses [79]. Another viable explanation is that the structure of the used indoor

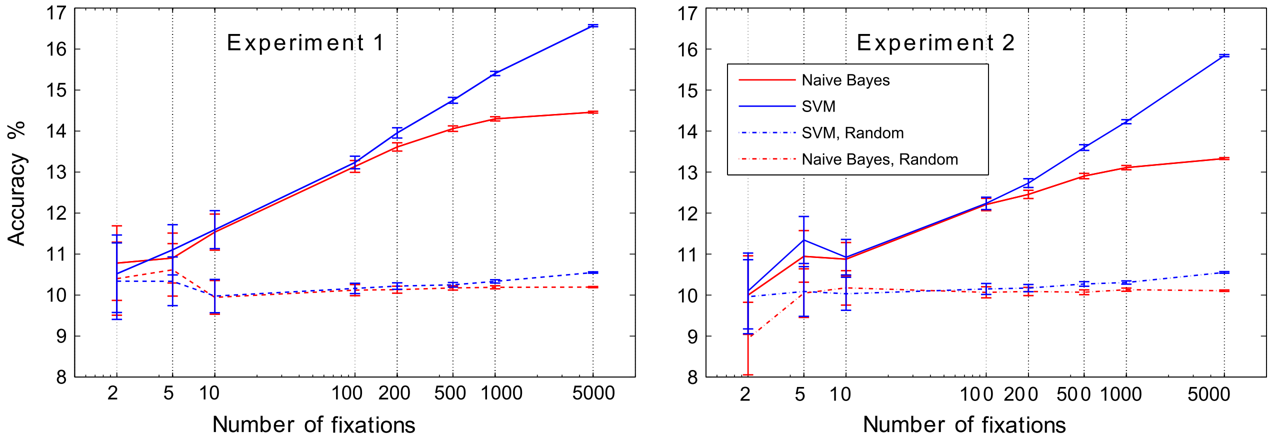

Fig. 6. Search target decoding accuracy for increased number of fixations using SVM and Naive Bayes classifiers in Experiments 1 and 2. Chance level here is at 10%. Error bars indicate standard error of the mean (s.e.m) across 200 runs.

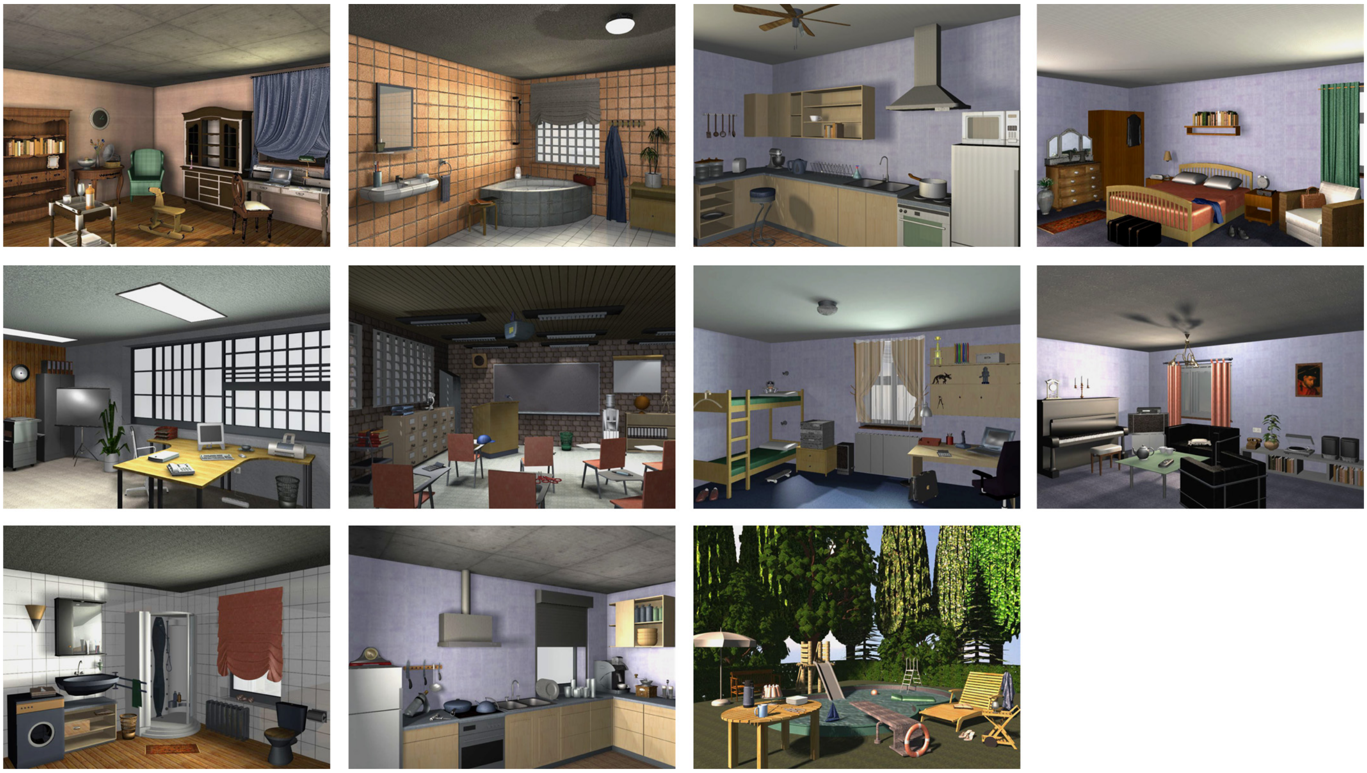

Fig. 7. Stimuli used in Experiment 3. Data was borrowed from [83] in which six subjects searched for one of 15 objects in 11 full-color, 3D rendered images of real-world scenes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

scene dataset is such that wherever one fixates, there is sufficient information to perform target inference.

The best performance of 25.4% was achieved using RGB Hist features, window size 129, and 15 fixations. This result is significantly above the chance level of 6.7% (t-test, $p < 0.01$) and trending towards significantly higher decoding accuracy than using fixation locations (t-test, $p = 0.1$). Averaged over all fixations (x-axis), SVM using RGB Hist results in 22.7% accuracy which is significantly above 21.7% using fixation locations (t-test, $p = 0.02$). Increasing the number of fixations to 25 did not change the results dramatically (maximal accuracy of 26% using RGB Hist).

As a control analysis, we trained a classifier with random fixations and randomly extracted patches (using RGB Hist). A number of random fixations matching that of human fixations were extracted in search for each object. Classification accuracy with random fixations rose with more fixations up to 15.15% at 15 fixations which is significantly lower than accuracy using human fixations (t-test, $p < 0.005$). Similarly, using random patches resulted in accuracy of 15.4% which is significantly lower than accuracy using human fixated patches ($p < 0.005$). As the number of random fixations increases, the algorithm gets to sample more of the scene, and therefore, the accuracy of target prediction increases.

5. Discussion

We reported algorithms for search target inference from eye movements in synthetic and natural scenes. The results provide insight into both the way humans search for complex patterns and the most promising approaches to extracting intent information from fixation data. With regard to human search processes, it seems that complex target patterns are not matched with local display areas as a whole but that search is guided by subpatterns (in alignment with findings from [68]). Only target verification processes use the complete memorized target information. Pattern voting takes advantage of this fact and thus is superior over target inference approaches based on classification images [68]. Furthermore, the focus of attention during real-world scene search appears to be matched in size with typical individual scene objects.

Here we treated all visited search patterns equally. While there does not seem to be an obvious reason why certain fixations during the search may convey more information than others regarding target appearance (e.g., fixations near the start of trial vs. fixations near the end of trial), a systematic investigation of this factor may tell us how subjects accumulate information over time during search (see for example [56]).

Note that the problem addressed here does not fit to the classic setting in the pattern recognition literature in which a set of data from all classes along with their labels are available to train classifiers (i.e., many classes here do not have samples). In other words, no other obvious algorithms seem to be feasible in our scenario beyond the ones we tested. While it is possible to design search tasks with fewer possible search targets (as in Exp. 3), the advantage of our first two experiments is that they challenge algorithms to their extreme and force them to take better advantage of the limited amount of available data.

Is it possible to obtain better accuracies than what we reported in our first two experiments? It is very hard to answer this question as several factors such as subjects' strategies over space and time, their uncertainty in landing saccades, and eye tracking error are involved. Even a human benchmark here would not be very useful and most likely humans will fail to report the search target from fixated patterns. Failure of humans, however, does not necessarily mean such information does not exist in the data. In a similar setting to ours, Greene et al., [28] asked a group of observers to watch the scanpaths of some other observers and guess the tasks under which they viewed images (i.e., Yarbus' experiment). They found that their observers were not able to

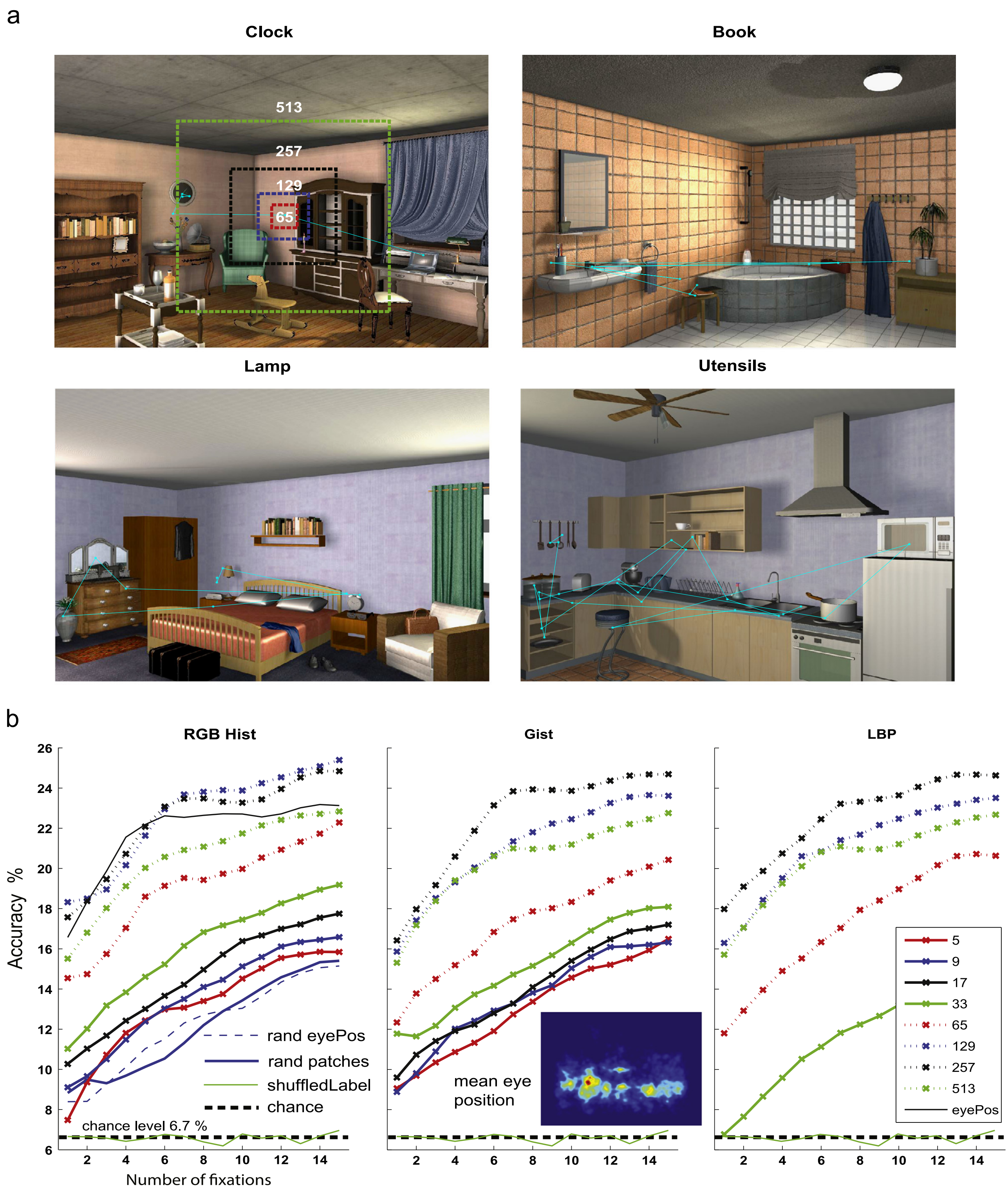

Fig. 8. (a) Search scanpaths for 4 objects in natural scenes. In alignment with the previous works (e.g., [27,14]), we qualitatively observe that viewers look more often at the objects (which are usually unstructured) rather than the background (which is usually structured). (b) Decoding accuracy using 3 feature types and multi-class SVM for increasing number of fixations and surrounding window sizes. Solid black curve is the accuracy of an SVM classifier with fixation position information only. Mean eye position over all search trials (inset) shows a bias toward lower scene regions as most objects often appear on the floor. Classification accuracies using all feature types are significantly above chance, indicating that fixated patches convey information regarding the search target. An SVM with shuffled labels (from [1 15]) results in 6.7% accuracy (chance-level).

perform better than chance. This (along with the failure of their classifiers trained from eye movement patterns) led them to conclude that scanpaths lack diagnostic information regarding the task. Interestingly, later studies [6,41] were able to classify observers' task from eye movement patterns indicating the presence of task diagnostic information in scanpaths.

In Experiment 3, we used data of Võ and Wolfe [83] and showed that search target inference is feasible on their data to

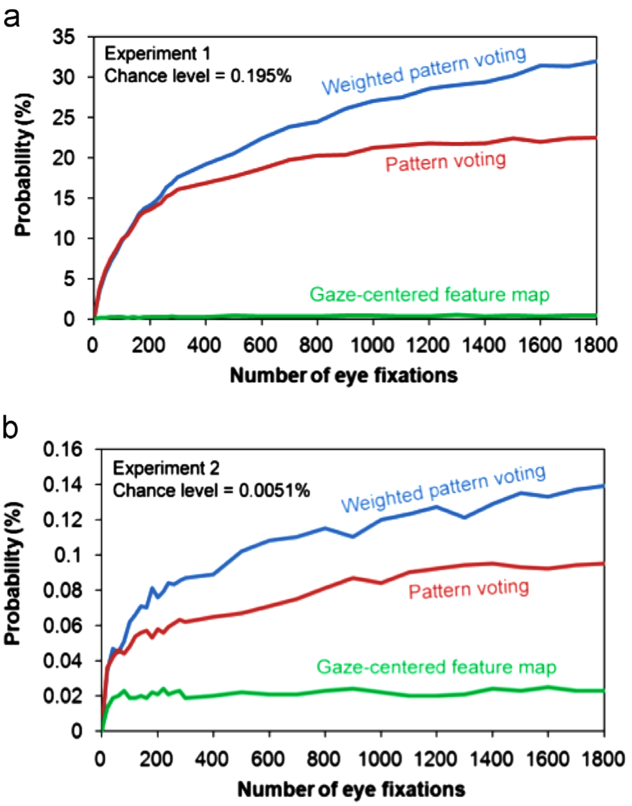

Fig. 9. Search target prediction accuracy when the target is among the patterns with the maximum vote (see Appendix A).

some extent using pattern classifiers. However, it is not trivial to readily apply our pattern voting algorithms to this data due to the high variability of image data and the number of possible target patterns (compared to our synthetic patterns). The bottom line is that target features offer more information than previously thought from Rajashekar et al.'s study [68] but demand more intelligent ways to extract and exploit them.

Despite a large volume of research on human behavior in visual search, to the best of our knowledge, no one has yet tried to perform the reverse which is predicting search targets in natural scenes from search statistics (here target features). We believe that our results together with the previous findings on the role of saliency and context in visual search (e.g., [26,79]) can help design more accurate search target inference algorithms in the future.

To move from target inference in visual search to complex intent and task decoding scenarios, we propose the following four categories of features. These features have been used scatteringly in the previous studies and include:

1. Oculomotor-based saccade metrics including distributions of saccade duration, inter-saccade interval, saccadic peak velocity, saccade amplitude, etc.
2. Stimulus-based features including distribution of features at fixated locations and spatial patterns of eye movements inspired by [85,24].
3. Correlations with bottom-up saliency maps (differential distributions of salience values at human gaze vs. random locations).
4. Correlations with top-down relevance maps. These top-down maps can be obtained by learning models for a specific task from training data. For example, a relevance map can be obtained by training a classifier from computer vision features to the task output y (e.g., W in a linear classifier Y=WX + b can be converted to a 2D saliency map). This relevance map is basically equivalent to annotating data and then measuring feature statistics on annotated regions (for features such as face, text, and gaze direction). The underlying idea here is to generate features that represent a fusion of image-derived information (salience/relevance maps) with observer-derived behavior (eye positions). This means fusing stimulus and behavior.

The above features can be used to train different types of algorithms and pattern classifiers for answering the following questions regarding stimulus, observer's identity, and his task:

1. Given the stimulus x and pattern of fixations e, what would be the observer's response y?
2. Given response y and patterns of eye movements e, what is stimulus x or its category? Such prediction is possible because different images have different informative or interesting regions that attract attention (e.g., a beach scene vs. a city scene).
3. Given e and x (or just e), which task, out of several tasks T^1, T^2, …, T^n, a subject has been performing (i.e., Yarbus's classic experiment [85])?
4. Given $(x, y$, and $e)$, what is the identity or category of the observer (e.g., healthy or patient [81])? This might be inferable due to idiosyncratic patterns of eye movements (e.g., [20]).
5. Given x and T, what are the locations of fixations e? This is the goal of gaze prediction and saliency models.

We found that, on synthetic search tasks, subjects break down complex target patterns to smaller manageable pieces to look for in a scene. This aligns with previous findings (e.g., [84,82,58,13,19]) that suggest neurons to be biased in a top-down manner to render the target more salient. Such studies, however, have mainly considered elementary features such as orientation, color, and intensity which are believed to be extracted by early visual areas. Our results extend these findings and suggest that biasing of more complex mid-level features or composition of simple features might be also possible.

6. Conclusion

The present data suggest that the mechanisms underlying the weighted pattern voting algorithm are robust enough for a useful target estimation in a variety of human–computer interfaces. Our target inference algorithms can be adapted to various display types, since image filters commonly used in computer vision and behavioral studies (e.g., [88]) can transform any display into a matrix of feature vectors. Moreover, the current data advocate that the future designers of smart, gaze-controlled human–computer interfaces should keep the spatial complexity of display objects low in order to induce more distinctive patterns of eye movements for individual search targets.

We also mentioned what features can be extracted from eye movement data for general intent decoding. These features can be used for developing algorithms that can be utilized for a number of applications such as wearable visual technologies (smart glasses like Google Glass), smart displays (phones and tablets), adaptive web search, activity recognition, human–computer interaction, biometrics, marketing, and patient diagnosis (e.g., Autistic, ADHD, Parkinson, Alzheimer). Eventually, eye movements together with EEG, fMRI, cell recording data, and purely biological cues such as pupil size, sweating, heart rate, and breathing can lead to high-throughput intent decoding.

Acknowledgments

A.B. was supported by the National Science Foundation (Grant number CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and U.S. Army (W81XWH-10-2-0076). M.P. was supported by Grant number R15EY017988 from the National Eye Institute. The authors would like to thank Melissa Le-Hoa Võ and Jeremy Wolf for sharing their data with us. Thanks also to reviewers for their valuable comments. Our code and data are publicly available at http://ilab.usc.edu/borji/Resources.html.

Appendix A

In Section 3, to break the tie in votes (i.e., when there are several patterns with the same maximum votes), we randomly selected one of the patterns with most votes. Here, we report results with an alternative criterion: a hit is declared when the target is among the patterns with the same maximum votes. Fig. 9 – the results for 1800 fixations. As it shows, patterns are the same as those shown in Fig. 4 but accuracies are about two times higher. The feature gaze map algorithm is executed here over only fixated patterns (and not their neighbors).

Appendix B

The weighted pattern voting algorithm was determined as the strongest target inference method through evaluation of various approaches. Besides the algorithms presented above, several other algorithms were implemented, their parameters and components were fitted using ten subjects' data, and they were evaluated on the other ten subjects data. The fitting of the algorithms included modifications for gaze-anchor estimation, which led to relative performance gains of up to 72.4% (using the criterion in Appendix A). The algorithms and their performance (for the strongest-performing variant and parameters) based on 1800 fixations in Experiments 1 and 2, respectively, were the following: pattern voting with votes weighted by similarity (30.9% and 0.122%), pattern voting weighted by fixation duration (24.1% and 0.115%), pattern voting with lower (7) similarity threshold (21.5% and 0.102%), Bayesian inference based on similarity metric (18.5% and 0.098%), 2×2 subpattern voting (9.2% and 0.115%), 3×1 and 1×3 subpattern voting (7.1% and 0.103%), feature map based on most frequently fixated patterns (6.5% and 0.08%), voting based on feature correlation between neighboring squares (6% and 0.093%), and average luminance in gaze-centered window (0.26% and 0.0069%).

References

[1] D. Ballard, M. Hayhoe, J. Pelz, Memory representations in natural tasks, J. Cognit. Neurosci. 7 (1) (1995) 66–80.
[2] T. Betz, T.C. Kietzmann, N. Wilming, P. König, Investigating task-dependent top-down effects on overt visual attention, J. Vis. 10 (15) (2010).
[3] A. Borji, Boosting bottom-up and top-down visual features for saliency estimation, in: CVPR, 2012.
[4] A. Borji, L. Itti, Exploiting local and global patch rarities for saliency detection, in: CVPR, 2012.
[5] A. Borji, L. Itti, State-of-the-art in modeling visual attention, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 35 (1) (2013) 185–207.
[6] A. Borji, L. Itti, Defending Yarbus: eye movements reveal observers' task, J. Vis. 14 (3) (2014) 29.
[7] A. Borji, D.N. Sihite, L. Itti, Computational modeling of top-down visual attention in interactive environments, in: Proceedings of the British Machine Vision Conference (BMVC), pages 85.1–85.12, 2011.
[8] A. Borji, D.N. Sihite, L. Itti, An object-based bayesian framework for top-down visual attention, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[9] A. Borji, D.N. Sihite, L. Itti, Probabilistic learning of task-specific visual attention, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[10] A. Borji, D.N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study, IEEE Trans. Image Process. 22 (1) (2012) 55–69.
[11] A. Borji, D.N. Sihite, L. Itti, What/Where to look next? Modeling top-down Visual Attention in Complex Interactive Environments, IEEE T. Syst. Man Cybern. Syst. 44 (5) (2014) 523–538.
[12] A. Borji, H.R. Tavakoli, D.N. Sihite, L. Itti, Analysis of scores, datasets, and models in visual saliency modeling, in: Proceedings of the International Conference on Computer Vision (ICCV), 2013.
[13] Ali Borji, Laurent Itti, Optimal attentional modulation of a neural population, Front. Comput. Neurosci. 8 (2014).
[14] Ali Borji, Dicky N. Sihite, Laurent Itti, What stands out in a scene? A study of human explicit saliency judgment, Vis. Res. 91 (2013) 62–77.
[15] A. Bulling, J.A. Ward, H. Gellersen, G. Tröster, Eye movement analysis for activity recognition using electrooculography, IEEE Trans. Pattern Anal. Mach. Intell. 33 (April 4) (2011) 741–753.
[16] G.T. Buswell, How People Look at Pictures, University of Chicago Press, Chicago, 1935.
[17] C. Corinna, V. Vladimir, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[18] R.C. Cagli, P. Coraggio, P. Napoletano, O. Schwartz, M. Ferraro, G. Boccignone, Visuomotor characterization of eye movements in a drawing task, Vis. Res. 49 (2009).
[19] M. Carrasco, Visual attention: the past 25 years, Vis. Res. 51 (2011) 1484–1525.
[20] M.S. Castelhano, M.L. Mack, J.M. Henderson, Viewing task influences eye movement control during active scene perception, J. Vis. 9 (3) (2009) 1–15.
[21] K. Chajka, M.M. Hayhoe, B. Sullivan, J. Pelz, N. Mennie, J. Droll, Predictive eye movements in squash, J. Vis. 6 (6) (2006) 481.
[22] J. Clark, J. O'Regan, Word ambiguity and the optimal viewing position in reading, Vis. Res. 4 (39) (1998) 843–857.
[23] S. Crespi, C. Robino, O. Silva, C. deSperati, Spotting expertise in the eyes: billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task, J. Vis. 12 (11) (2012).
[24] M. DeAngelus, J.B Pelz, Top-down control of eye movements: Yarbus revisited, Vis. Cognit. 17 (6-7) (2009) 790–811.
[25] A. Doshi, M.M. Trivedi, On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes, IEEE Trans. Intell. Transp. Syst. 10 (3) (2009) 453–462.
[26] K.A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, A. Oliva, Modelling search for people in 900 scenes: a combined source model of eye guidance, Vis. Cognit. 17 (6–7) (2009) 945–978.
[27] Lior Elazary, Laurent Itti, Interesting objects are visually salient, J. Vis. 8 (January (3)) (2008), pp. 3.1–15.
[28] M.R. Greene, T. Liu, J.M. Wolfe, Reconsidering Yarbus: a failure to predict observers' task from eye movement patterns, Vis. Res. 62 (2012) 1–8.
[29] N. Hagemann, J. Schorer, R. Cañal-Bruland, S. Lotz, B. Strauss, Visual perception in fencing: do the eye movements of fencers represent their information pickup? Atten., Percept., Psychophys. 72 (8) (2010) 2204–2214.
[30] A. Haji-Abolhassani, J.J. Clark, A computational model for task inference in visual search, J. Vis. 29 (13) (2013) 1–24.
[31] J. Harel, C. Moran, A. Huth, W. Einhäuser, C. Koch, Decoding what people see from where they look: predicting visual stimuli from scanpaths, Lecture Notes in Artificial Intelligence (WAPCV), 2008.
[32] Mary M. Hayhoe, Anurag Shrivastava, Ryan Mruczek, Jeff B. Pelz, Visual memory and motor planning in a natural task, J. Vis. 3 (1) (2003).
[33] M.M. Hayhoe, D.H. Ballard, Eye movements in natural behavior, Trends Cogn. Sci. 9 (April (4)) (2005) 188–194.
[34] J.M. Henderson, Human gaze control during real-world scene perception, Trends Cogn. Sci. 7 (11) (2003) 498–504.
[35] J.M. Henderson, S.V. Shinkareva, J. Wang, S.G. Luke, J. Olejarczyk, Predicting cognitive state from eye movements, Plos One 8 (2013).
[36] A.J. Hornof, Cognitive strategies for the visual search of hierarchical computer displays, Human.–Comput. Interact. 19 (2004) 183–223.
[37] A.D. Hwang, H.C. Wang, M. Pomplun, Semantic guidance of eye movements in real-world scenes, Vis. Res. 51 (2011) 1192–1205.
[38] S.T. Iqbal, B.P. Bailey, Using eye gaze patterns to identify user tasks, in: The Grace Hopper Celebration of Women in Computing, 2004.
[39] L. Itti, C. Koch, Computational modelling of visual attention, Nat. Rev. Neurosci. 2(3) (2001) 194–203.
[40] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (November (11)) (1998) 1254–1259.
[41] C. Kanan, N. Ray, D.N. Bseiso, J.H. Hsiao, G.W. Cottrell, Predicting an observer's task using multi-fixation pattern analysis, in: Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA), 2014.
[42] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (4) (1985) 219–227.
[43] G. Kuhn, B.W. Tatler, Magic and fixation: now you dont see it, now you do, Perception 34 (9) (2005) 1155–1161.
[44] M. Land, N. Mennie, J. Rusted, The roles of vision and eye movements in the control of activities of daily living, Perception 28 (11) (1999) 1311–1328.
[45] M.F. Land, D.N. Lee, Where we look when we steer, Nature 369 (1994) 742–744.

[46] M.F. Land, P. McLeod, From eye movements to actions: how batsmen hit the ball, Nat. Neurosci. 3 (December (12)) (2000) 1340–1345.

[47] M.F. Land, Eye movements and the control of actions in everyday life, Prog. Retin. Eye Res. 25 (2006) 296–324.

[48] M.F. Land, M.M. Hayhoe, In what ways do eye movements contribute to everyday activities? Vis. Res. 41 (25–26) (2001) 3559–3565.

[49] G.E. Legge, T.S. Klitz, B.S. Tjan, Mr. Chips: an ideal-observer model of reading, Psychol. Rev. 104 (3) (1997).

[50] F. Lethaus, M.R. Baumann, F. Köster, K. Lemmer, A comparison of selected simple supervised learning algorithms to predict driver intent based on gaze data, Neurocomputing 121 (2013) 108–130.

[51] T. Loetscher, C.J. Bockisch, M.E. Nicholls, P. Brugger, Eye position predicts what number you have in mind, Curr. Biol. 20 (2010) 264–265.

[52] N.A. Macmillan, C.D. Creelman, Detection Theory: A Users Guide, 1991.

[53] F.W. Mast, S.M. Kosslyn, Eye movements during visual mental imagery, Trends Cogn. Sci. 6 (2002).

[54] N. Mennie, M. Hayhoe, B. Sullivan, Look-ahead fixations: anticipatory eye movements in natural tasks, Exp. Brain Res. 179 (3) (2007) 427–442.

[55] R.C. Miall, J. Tchalenko, A painter's eye movements: a study of eye and hand movement during portrait drawing, Leonardo: Int. J. Arts Sci. 34 (2001) 35–40.

[56] J. Najemnik, S.W. Geisler, Optimal eye movement strategies in visual search, Nature 434 (7031) (2005) 387–391.

[57] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vis. Res. 45 (2005) 205–231.

[58] V. Navalpakkam, L. Itti, Search goal tunes visual features optimally, Neuron 2/15 (2007).

[59] V. Navalpakkam, C. Koch, A. Rangel, P. Perona, Optimal reward harvesting in complex perceptual environments, Proc. Natl. Acad. Sci. 107 (2010).

[60] T. O'Connell, D. Walther, Fixation patterns predict scene category, in: VSS, 2012.

[61] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, Pattern Anal. Mach. Intel. IEEE Trans. 24 (7) (2002) 971–987.

[62] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.

[63] J.P. Ossandón, S. Onat, P. König, Spatial biases in viewing behavior, J. Vis. 14 (2) (2014) 20.

[64] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vis. Res. 42 (1) (2002) 107–123.

[65] R.J. Peters, L. Itti, Congruence between model and human attention reveals unique signatures of critical visual events, in: Advances in Neural Information Processing Systems (NIPS), 2007.

[66] M. Pomplun, Saccadic selectivity in complex visual search displays, Vis. Res. 46 (12) (2006) 1886–1900.

[67] Marc Pomplun, Tyler W. Garaas, Marisa Carrasco, The effects of task difficulty on visual search strategy in virtual 3d displays, J. Vis. 13 (3) (2013) 24.

[68] J. Rajashekar, L.C. Bovik, A.K. Cormack, Visual search in noise: revealing the influence of structural cues by gaze-contingent classification image analysis, J. Vis. 17 (2006) 379–386.

[69] K. Rayner, Eye guidance in reading: fixation locations within words, Perception 8 (1979).

[70] E.D. Reichle, K. Rayner, A. Pollatsek, The E–Z reader model of eye movement control in reading: comparisons to other models, Behav. Brain Sci. 26 (2003) 445–476.

[71] L.W. Renninger, J.M. Coughlan, P. Verghese, J. Malik, An information maximization model of eye movements, in: Advances in Neural Information Processing Systems (NIPS), 2005.

[72] R.D. Rimey, C.M. Brown, Control of selective perception using Bayes nets and decision theory, Int. J. Comput. Vis. 12 (2/3) (1994) 173–207.

[73] R. Rosenholtz, Y. Li, L. Nakano, Measuring visual clutter, J. Vis. 7 (2) (2007).

[74] A.C. Schütz, D.I. Braun, K.R. Gegenfurtner, Eye movements and perception: a selective review, J. Vis. 11 (5) (2011) 1–30.

[75] S. Sears, J.A. Jacko, The Human–Computer Interaction Handbook, CRC Press, Lincoln, 2003.

[76] J. Shen, E.M. Reingold, Saccadic selectivity during visual search: the effects of shape and stimulus familiarity, in: Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society, 1999.

[77] N. Sprague, D. Ballard, Eye movements for reward maximization, In: Proceeding of Advances in neural information processing systems, 2003.

[78] B.W. Tatler, M.M. Hayhoe, M.F. Land, D.H. Ballard, Eye guidance in natural vision: reinterpreting salience, J. Vis. 11 (5) (2011) 1–23.

[79] A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, Psychol. Rev. 113 (4) (2006) 766.

[80] A.M. Treisman, G. Gelade, A feature integration theory of attention, Cogn. Psychol. 12 (1980) 97–136.

[81] P. Tseng, I.G.M. Cameron, G. Pari, J.N. Reynolds, D.P. Munoz, L. Itti, High-throughput classification of clinical populations from natural viewing eye movements, J. Neurol. (2012).

[82] Preeti Verghese, Visual search and attention: a signal detection theory approach, Neuron 31 (4) (2001) 523–535.

[83] M.L.H. Võ, J.M. Wolfe, The interplay of episodic and semantic memory in guiding repeated search in scenes, Cognition 126 (2013) 198–212.

[84] J.M. Wolfe, Visual search, in: H. Pashler (Ed.), Attention, Psychology Press, Hove, UK, 1998, pp. 13–71.

[85] A.L. Yarbus, Eye movements and Vision, Plenum, New York, 1967.

[86] W. Yi, D. Ballard, Recognizing behavior in hand-eye coordination patterns, Int. J. Humanoid Robot. 6 (03) (2009) 337–359.

[87] G. Zelinsky, W. Zhang, D. Samaras, Eye can read your mind: decoding eye movements to reveal the targets of categorical search tasks [abstract], J. Vis. 8 (2008) 380.

[88] G.J. Zelinsky, A theory of eye movements during target acquisition, Psychol. Rev. 115 (October (4)) (2008) 787–835.

Ali Borji received the B.S. and M.S. degrees in computer engineering from the Petroleum University of Technology, Tehran, Iran, 2001 and Shiraz University, Shiraz, Iran, 2004, respectively. He received the Ph.D. degree in computational neurosciences from the Institute for Studies in Fundamental Sciences (IPM) in Tehran, 2009. He then spent a year at the University of Bonn as a postdoc. He has been a postdoctoral scholar at iLab, University of Southern California, Los Angeles, since March 2010. His research interests include computer vision, machine learning, and neurosciences with particular emphasis on visual attention, visual search, active learning, scene and object recognition, and biologically plausible vision models.

Andreas Lennartz worked as a visiting masters student at the Department of Computer Science, University of Massachusetts at Boston in 2006–2007. He is currently with System development team DWH at Air Berlin PLC & Co. KG aviation.

Marc Pomplun is an Associate Professor of Computer Science at the University of Massachusetts Boston, where he joined the faculty in 2003. He received his M.S. in Computer Science at Bielefeld University, Bielefeld, Germany, in 1994. He obtained his Ph.D. in Computer Science under Professor Dr. Helge Ritter at Bielefeld University, Bielefeld, Germany. He was a postdoctoral fellow in the Department of Psychology at the University of Toronto. His research interests include human vision, computer vision and Human–computer interaction.