The feasibility of gaze tracking for “mind reading” during search

Andreas Lennartz & Marc Pomplun

Department of Computer Science, University of Massachusetts at Boston,
100 Morrissey Blvd., Boston, MA 02125, USA

We perform thousands of visual searches[1,2] every day, for example, when selecting items in a grocery store or when looking for a specific icon in a computer display[3]. During search, our attention and gaze are guided toward visual features similar to those in the search target[4-6]. This guidance makes it possible to infer information about the target from a searcher’s eye movements. The availability of compelling inferential algorithms could initiate a new generation of smart, gaze-controlled interfaces that deduce from their users’ eye movements the visual information for which they are looking. Here we address 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? While we choose a random-dot search paradigm for these analyses to eliminate contextual influences on search[7], the proposed techniques can be applied to the local feature vectors of any type of display. We present an algorithm that correctly infers the target pattern up to 66 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 patterns exceed a certain spatial complexity level, only a subpattern tends to guide the observers' eye movements, which drastically impairs target inference.

Eye movements can reveal a wealth of information about the content of a person's visual consciousness, such as the current interpretation of an ambiguous figure[8] or the geometry of mental imagery[9]. During visual search, our eye movements are attracted by visual information resembling the search target[4-6], causing the image statistics near our fixated positions to be systematically influenced by the shape[10] and basic visual features[5] of the target. One study found that the type of object sought, of two possible categories, can be inferred from such systematic effects [11]. If such inference were possible for a larger set of candidate objects, a new generation of smart, gaze-controlled human-computer interfaces[12] could become reality. 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 users.

To explore the potential of algorithmically inferring the search target from a searcher's eye fixations, we conducted two experiments of visual search in random-dot patterns (see Fig. 1). 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 was excluded, because the disproportionate fixation density at the end of a search would have made target inference trivial. A variety of inferential algorithms (see Methods) was 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 for a first quantitative exploration of this field, because it minimized the influence of semantic factors[7] on eye movements and supplied fixed numbers of equally probable target patterns, $2^9 = 512$ in Experiment 1 and $3^9 = 19683$ 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.

Our development and evaluation of inferential algorithms resulted in the discovery of two particularly powerful mechanisms, whose combination outperformed all other methods for both Experiments 1 and 2 without modifying its parameters between experiments. In the following, we will describe these two components and compare their performance with an approach adapted from a previous study[10]. In that study[10], 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, for every fixation, we extracted the display data from a 3×3-square window whose center square was placed over the square on which the fixation landed. We computed the frequency of each feature (black, gray, and white) in each square and subtracted the average frequency of that feature across the nine squares. The feature with the highest value in each square entered the estimated target pattern. This algorithm, which we termed 'gaze-centered feature map,' outperformed all other methods analyzing feature statistics in individual squares relative to fixation.

Our first newly developed technique, 'pattern voting,' is based on the assumption, derived from a previous study[6], 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. The voting algorithm keeps score of the votes for every possible 3×3 pattern. For each fixated square, a 3×3 window is placed over it nine times so that each of its squares lands on the fixated square once. Each time, the patterns whose similarity to the 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.

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. Our analyses showed that the most effective weights are obtained 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:

$$s_n = \sum_{r=1}^{R} \left(\frac{N \cdot v_{r,n}}{V_r} \right)^{\ln\left(e + \frac{V_r}{c}\right)} \quad \text{for } n = 1, \ldots, N, \qquad (1)$$

where N is the total number of patterns (512 or 19683 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 patterns in *r*, and *c* is a constant whose optimal value was found near 600 for both current experiments.

To evaluate the performance of the algorithms as a function of the number of available search fixations, we resampled those fixation datasets that were not used for developing the algorithms, that is, we repeatedly selected random subsets of them. Fig. 2 illustrates that pattern voting clearly outperforms the gaze-centered feature map. In Experiment 1, even after only 20 fixations (about 5 s of search), the voting algorithm's probability of picking the correct target is already 18.6 times above chance level, while it is only 0.2 times above chance for the feature map. After approximately 180 fixations, the weighted pattern voting starts surpassing the basic voting algorithm and maintains a steeper increase until the final 1800 fixations, where its performance reaches 31.9%, outperforming the voting algorithm (22.5%), $p < 0.01$, which in turn exceeded the performance of the gaze-centered feature map (0.5%), $p < 0.001$. This pattern is similar in Experiment 2 (0.139%, 0.095%, and 0.023%, respectively, for 1800 fixations, both *p*s < 0.05) but the superiority of the pattern voting algorithms over the feature map approach is less pronounced. Fig. 3 illustrates the top ranking choices made by the weighted pattern voting algorithm.

Even if we compensate for the difference in pattern set size, weighted pattern voting still performs clearly better in Experiment 1 than in Experiment 2, as indicated by greater performance-to-chance level proportion (163.5 versus 27.4, respectively), and sensitivity d' (2.54 versus 0.92, respectively) according to signal detection theory[13] (see Methods) , $p < 0.01$, for 1800 fixations. If the reason for this discrepancy were 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 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, and whenever this part is detected, a complete verification of the local pattern is conducted. To test this hypothesis, we used resampling (1800 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 (see Fig. 4a) was the top-ranked one. While the absolute hit rate does not differ statistically between Experiments 1 and 2 (68.1% versus 51.6%, respectively), $p > 0.3$, both the hit rate-to-chance level proportion (2.72 versus 10.44, respectively) and sensitivity d' (0.65 versus 1.19, respectively), are greater in Experiment 2, $p < 0.01$, supporting our hypothesis (Fig. 4b).

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. The proposed mechanisms can be adapted to various display types, since image filters commonly used in computer vision[14] and behavioral studies[5,15] 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.

Methods

Experiments. The same twenty subjects, aged 19 to 36 and having normal or corrected-to-normal vision, participated in each experiment after giving informed consent. Their eye movements were measured using an EyeLink-II head mounted eye tracker (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 cd/m^2) or white (71.2 cd/m^2) . In Experiment 2, a third luminance level (gray, 36.2 cd/m^2) was added. Subjects had six seconds 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 (see 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 30 trials during which the search target remained the same.

Algorithms. Several other algorithms were implemented, their parameters and components fitted – including modifications for gaze-anchor estimation leading to relative performance gains of up to 72% - using ten subjects' data and evaluated on the other ten subjects' data. These algorithms and their performance based on 1800 fixations in Experiments 1 and 2, respectively, were: 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%).

Sensitivity computation. 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[13,16]. 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[13]. In the subpattern analysis (Fig. 4), we further make the simplifying assumption that all subpatterns of a target are fixated with the same probability.

1. Wolfe, J.M. in *Attention*, H. Pashler, Ed. (Psychology Press, Hove, UK, 1998), pp. 13-71.

2. Najemnik, J. & Geisler, W.S. *Nature* **434**, 387-391 (2005).

3. Hornof, A. J. *Human-Computer Interaction* **19**, 183-223 (2004).

4. Wolfe, J.M. *Psychon. Bull. Rev.* **1**, 202-238 (1994).

5. Pomplun, M. *Vision Res.* **46**, 1886-1900 (2006).

6. Shen, J., & Reingold, E. M. In *Proc. of the Twenty-First Annual Conf. of the Cog. Sci. Soc.*, M. Hahn & S. C. Stoness, Eds.(Erlbaum, Mahwah, NJ. 1999), pp. 649-652.

7. Henderson, J.M. & Hollingworth, A. *Ann. Rev. Psych.* **50**, 243-271 (1999).

8. Pomplun, M., Velichkovsky, B.M. & Ritter, H. *Perception* **25**, 931-948 (1996)

9. Mast. F.W. & Kosslyn, S.M. *Trends in Cog. Sci.***6**, 271-272.

10. Rajashekar J., Bovik, L.C. & Cormack, A.K., *J. of Vision* **6**, 379–386 (2006).

11. Zelinsky, G., Zhang, W., & Samaras, D. [Abstract]. *J. of Vision* **8**, 380.

12. Sears, S. & Jacko, J.A. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (CRC Press, Lincoln, USA, 2003).

13. Macmillan, N.A. & Creelman, C.D. *Detection Theory: A User's Guide* (Cambridge University Press, New York, 1991).

14. Gonzalez, R.E. & Woods, R.C. *Digital Image Processing* (Prentice Hall, Upper Saddle River, 2002).

15. Zelinsky, G. J. *Psych. Rev.* **115**, 787-835 (2008).

16. Hacker, MJ, & Ratcliff, R. *Perception & Psychophysics* **26**, 168-170 (1979).

Acknowledgement. The project was supported by Grant Number R15EY017988 from the National Eye Institute to M.P.

Correspondence and requests for materials should be addressed to M.P. (marc@cs.umb.edu).

Fig. 1. Search targets (left) and cut-outs from corresponding visual search displays with a human subject's scanpath superimposed on it. 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. **a**, Experiment 1; **b**, Experiment 2.

Fig. 2. Comparison of inferential performance of the gaze-centered feature map algorithm adapted from a related study[10] and the two pattern voting algorithms proposed in the current study. Performance is measured as the probability of correctly inferred target objects as a function of the number of gaze fixations provided to the algorithms. This probability was approximated by repeated resampling (20,000 and 100,000 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 19683 in Experiment 2. **a**, Experiment 1; **b**, Experiment 2.

Fig 3. Actual targets (green frame) and the three patterns ranked highest by the weighted pattern voting algorithm in Experiment 1 (left) and Experiment 2 (right) based on all recorded fixations. 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 19683 candidates).

Fig. 4. Analysis of target subpattern frequencies near fixation. **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.

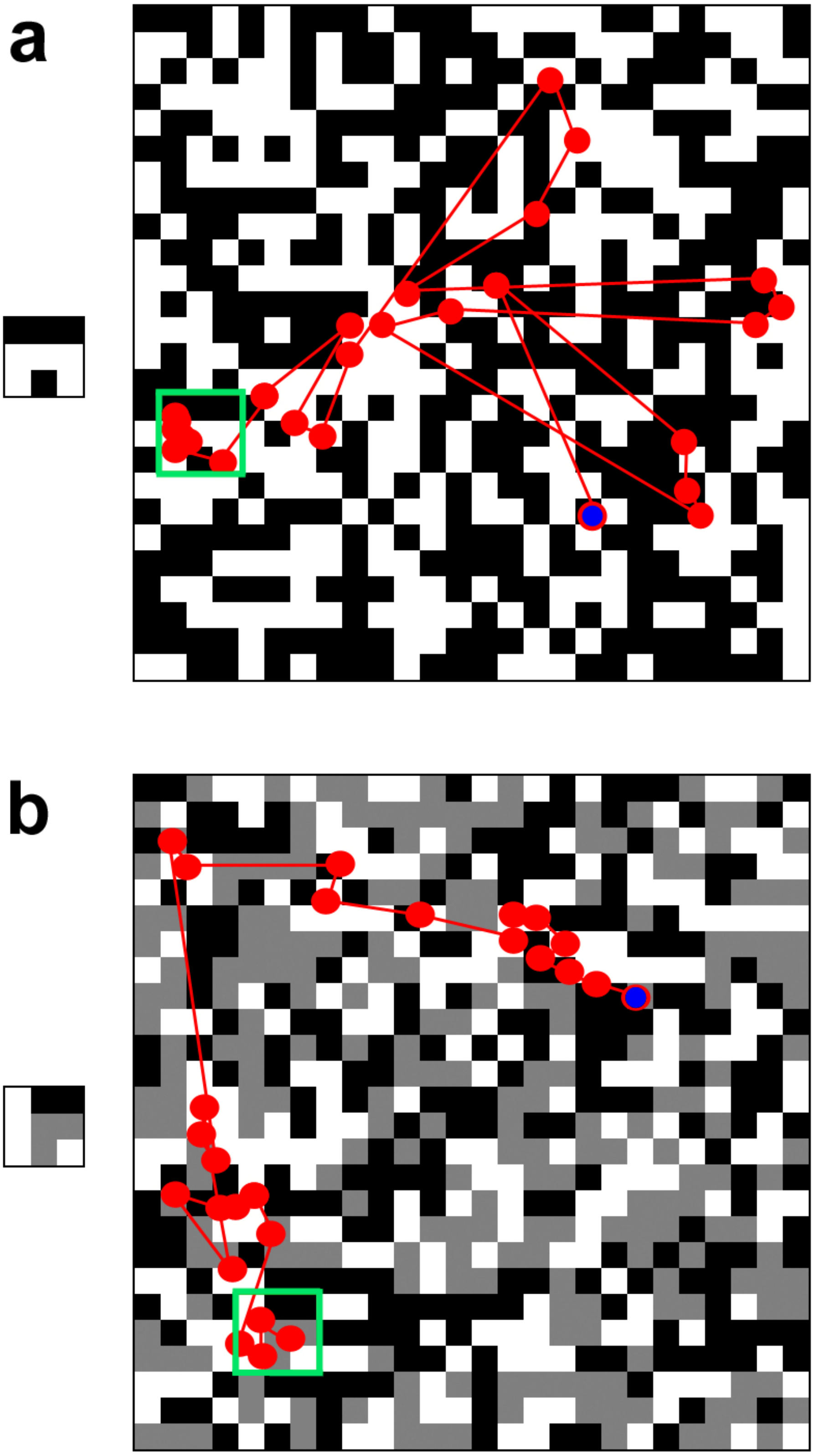

Figure 1

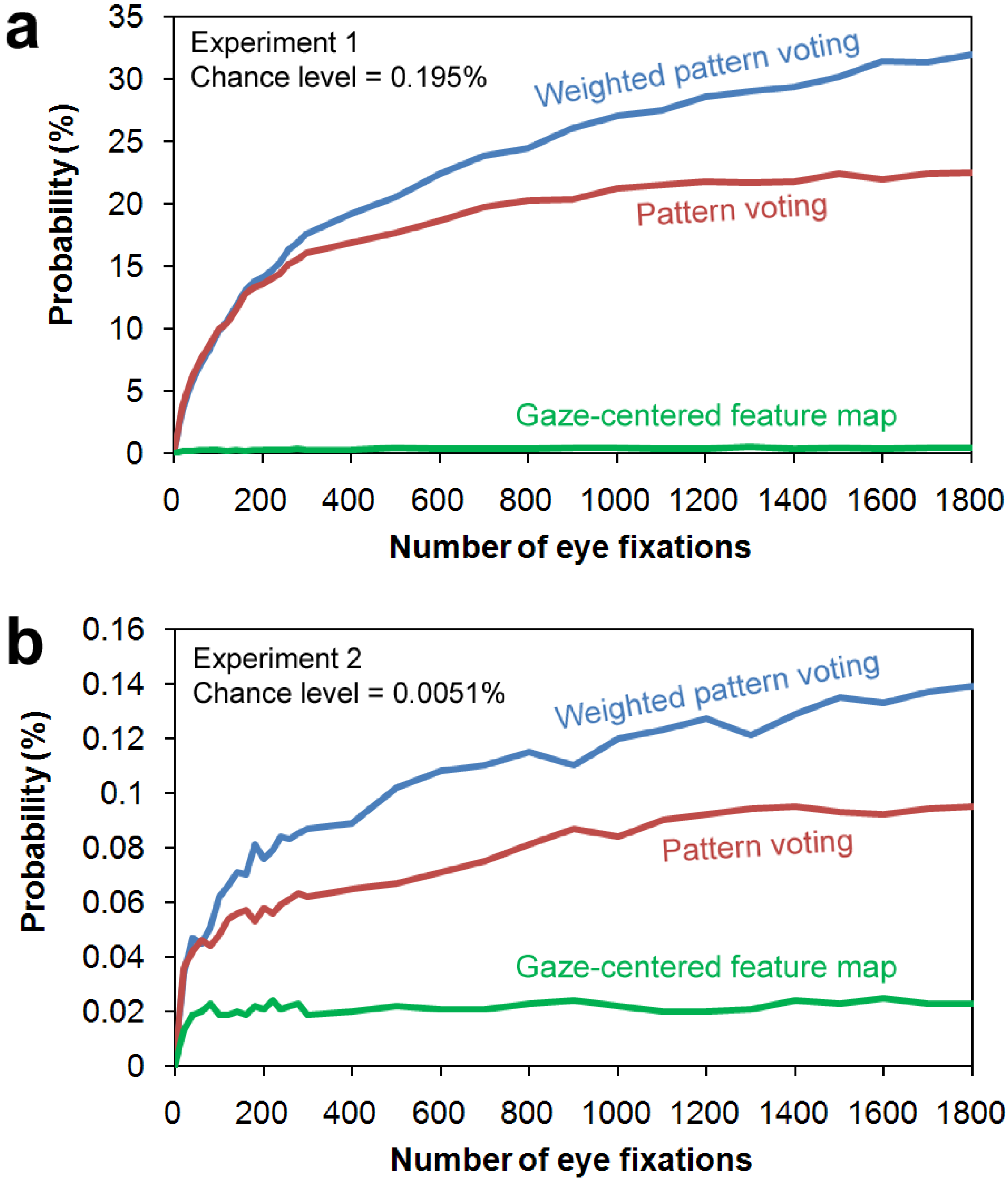
a
Experiment 1
Chance level = 0.195%
Weighted pattern voting
Pattern voting
Gaze-centered feature map
Probability (%)
35
30
25
20
15
10
5
0
0
200
400
600
800
1000
1200
1400
1600
1800
Number of eye fixations
b
Experiment 2
Chance level = 0.0051%
Weighted pattern voting
Pattern voting
Gaze-centered feature map
Probability (%)
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
200
400
600
800
1000
1200
1400
1600
1800
Number of eye fixations

Figure 2

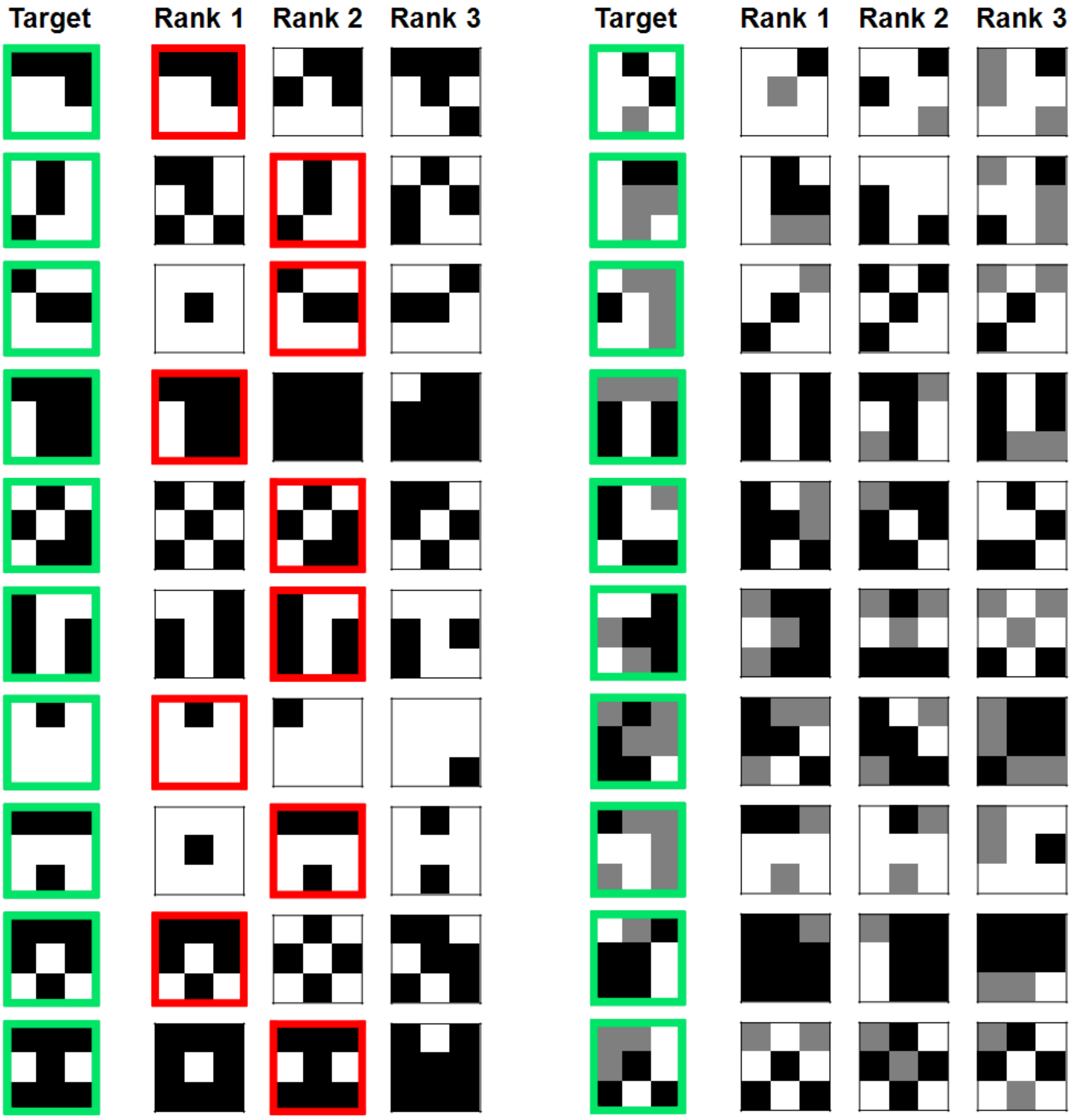

Figure 3

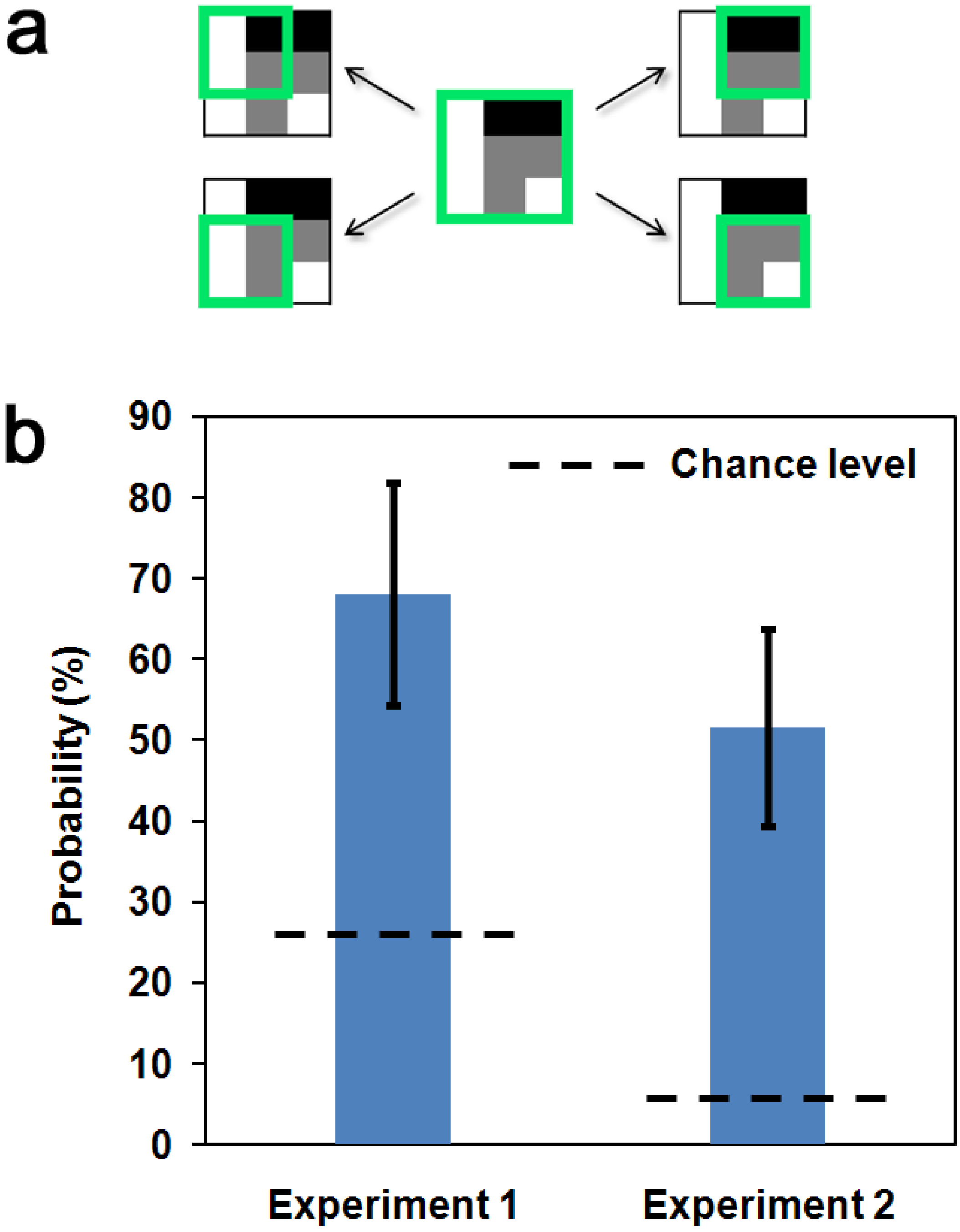
a
b
90
80
70
60
50
40
30
20
10
0
Probability (%)
Chance level
Experiment 1
Experiment 2

Figure 4