Active sensing in the categorization of visual patterns

Scott Cheng-Hsin Yang, Máté Lengyel, and Daniel M. Wolpert
Computational and Biological Learning Lab, Department of Engineering, University of Cambridge

The information that can be extracted from a visual scene depends on the sequence of eye movements chosen to scan the scene. Passive precomputed scan paths are suboptimal as the optimal eye movement will depend on past information gathered about the actual scene and prior knowledge about scene statistics and categories. Previous studies have identified active sensing in tasks in which the variable about which information needs to be sought is defined in the image space itself, such as the spatial location of a target in a scene during simple visual search (Najemnik & Geisler 2005) or when memorizing a shape (Renninger et al. 2007). However, it is still unknown whether eye movements are optimized to gather information about more abstract features of the visual scene such as visual categories.

We examined eye movement in a visual categorization task. Fur-like images were generated of three types: patchy, horizontal stripy, and vertical stripy (Fig. 1a). Participants had to categorize each image pattern as patchy or stripy (disregarding whether a stripy image was horizontal or vertical). The images were generated by Gaussian processes so that the individual pixel values varied widely even within a type and only higher order statistical information (ie. spatial correlation scales) could be used for categorization. We first presented the participants examples of full images to familiarize them with the statistics of the image. We then used a gaze-contingent display in which the entire pattern was initially occluded by a black mask and the underlying image was revealed with a small aperture at each fixation location (Fig. 1b). This allowed us to control the visual information available on each fixation so that the display would give the impression of viewing fur partially occluded by foliage.

We developed a Bayesian active sensor (BAS) algorithm for the task that uses knowledge of the statistics of the different visual patterns and the evidence accumulated from previous saccades to compute the probability of each pattern category and, hence, where to look next to achieve the maximal reduction in categorization error. We constructed an ideal observer which computed a posterior distribution, $P(c|D)$, over image category c given the data D (collection of previous revealing locations and revealed pixel values in the trial) and knowledge of the length scales corresponding to the different image types. The aim of BAS is to choose the next fixation location, x^*, so as to maximally reduce uncertainty in the category. This objective is formalized by the BAS scoring function which expresses the expected information gain when choosing x^*, and which can be conveniently computed as:

$$\mathrm{Score}(x^*|D) = \mathrm{H}[z^*|x^*, D] - \langle \mathrm{H}[z^*|x^*, D, c] \rangle_{p(c|D)}$$

where H denotes entropy (a measure of uncertainty), z^* is the possible pixel value at x^* and $\langle \cdot \rangle$ denotes averaging over the two categories weighted by their posterior probabilities (subscript). This expresses a trade-off between two terms. The first term encourages the selection of locations, x^*, where we have the most overall uncertainty about the pixel value, z^*, while the second term prefers locations for which our expected pixel value for each category is highly certain.

We first examined whether our participants used an active eye movement strategy. We show that their eye movement patterns, that is the density of fixation locations, depend on the underlying image patterns (Fig 2a first four rows and 2b), which could not be the case if a passive sensing strategy was used. To assess whether their active strategy contributed to performance improvement, we examined the same participants in a passive revealing condition. When the revealings were drawn randomly from an isotropic Gaussian centered on the image, performance was substantially impaired; however, when revealings were generated by noiseless BAS, performance improved substantially (Fig. 3a). This may seem to suggest that participants employed an active but suboptimal strategy to select their fixation locations, or, alternatively, this may be due to more trivial factors upstream or downstream of the process responsible for selecting the next fixation, such as noise and variability in perception or execution, respectively. To address this issue, we used the ideal observer model to quantify the amount of information accumulated about image category over subsequent revealings in a trial for different revealing strategies. Importantly, when we computed the information gain provided by BAS when operating with participants' perceptual noise, obtained by fitting our ideal observer model to their category choices, and typical saccadic variability reported in the literature, the discrepancy between the informativeness of BAS-generated revealings compared to participants disappeared (Fig. 3b). This suggests that the central component of choosing where to fixate was near-optimal in our participants, and suboptimality arose due to peripheral processes. Furthermore, eye movement patterns derived by BAS for the same images shown to our participants closely matched those in our participants: they were positively correlated with participants' eye movements for the same image type, but negatively correlated with those for different image types (Fig. 2a last row and 2c).

In conclusion, using our novel task analyzed in the framework of a Bayesian active sensor algorithm, we were able to show that participants were near-optimal in selecting each individual fixation location, so as to maximize information about image category, with performance only limited by low-level perceptual and motor variability.

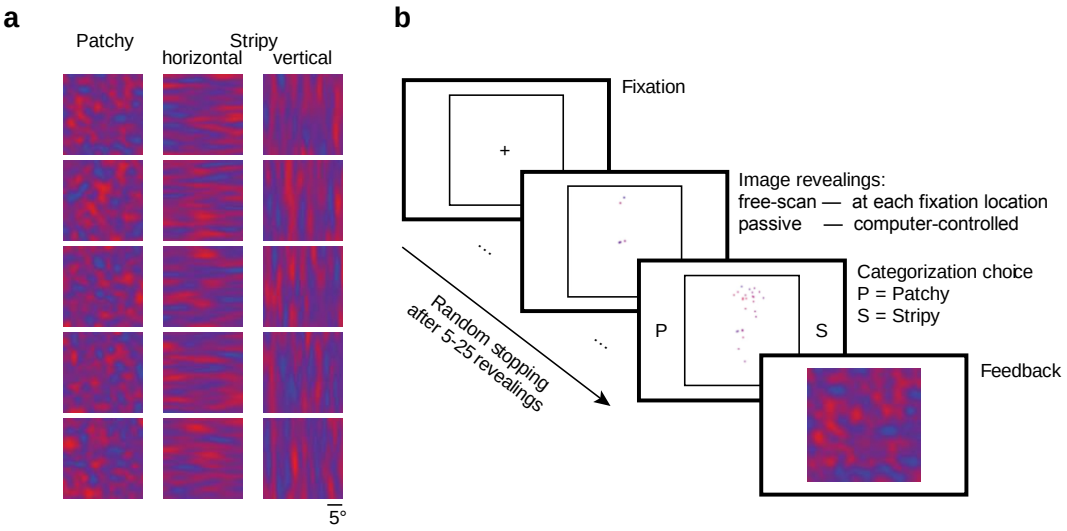

Fig. 1 Image categorization task. **a.** Example stimuli for each of the three image types sampled from two-dimensional Gaussian processes. **b.** Experimental design. Participants started each trial by fixating the center cross. In the free-scan condition, an aperture of the underlying image was revealed at each fixation location. In the passive condition, revealing locations were chosen by the computer. In both conditions, after a random number of revealings, participants were required to make a category choice (P vs. S) and were given feedback.

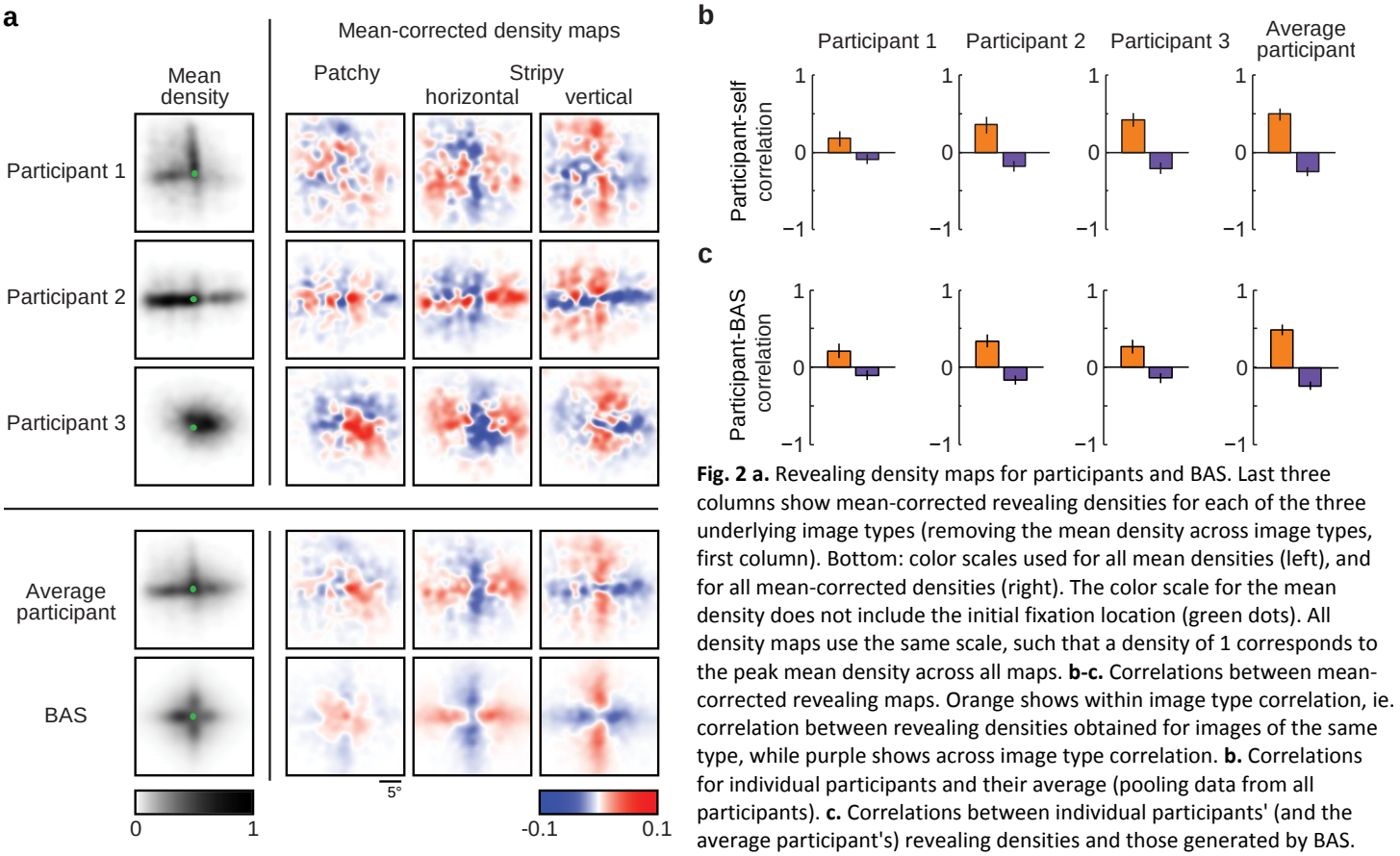

Fig. 2 a. Revealing density maps for participants and BAS. Last three columns show mean-corrected revealing densities for each of the three underlying image types (removing the mean density across image types, first column). Bottom: color scales used for all mean densities (left), and for all mean-corrected densities (right). The color scale for the mean density does not include the initial fixation location (green dots). All density maps use the same scale, such that a density of 1 corresponds to the peak mean density across all maps. **b-c.** Correlations between mean-corrected revealing maps. Orange shows within image type correlation, ie. correlation between revealing densities obtained for images of the same type, while purple shows across image type correlation. **b.** Correlations for individual participants and their average (pooling data from all participants). **c.** Correlations between individual participants' (and the average participant's) revealing densities and those generated by BAS.

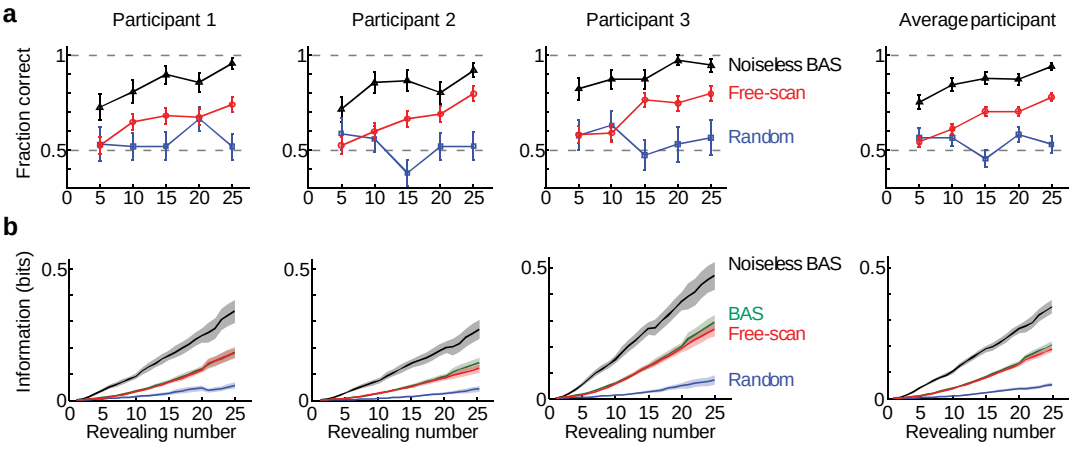

Fig. 3 Participants' performance in the task. **a.** Categorization performance as a function of revealing number for each of the three participants, and their average, under the free-scan and passive conditions corresponding to different revealing strategies. Error bars show s.e.m. across trials. **b.** Cumulative information gain of an ideal observer with different revealing strategies. Error bars show s.e.m. across trials.