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/

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

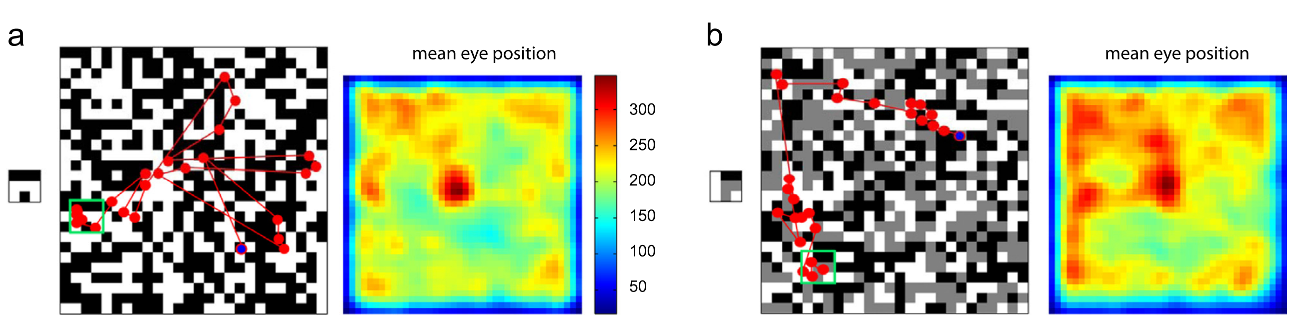

Fig. 1. Search targets and cut-outs from the corresponding visual search displays in (a) Experiment 1 and (b) Experiment 2 with human subjects' scanpaths superimposed on them. Actual displays consisted of 40×40 squares. Red discs indicate fixation positions, consecutive fixations are connected by straight 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.)