Advances in Cognitive Systems 4 (2016) 113–129 Submitted 8/2015; published 6/2016

Looking Around the Mind's Eye: Attention-Based Access to Visual Search Templates in Working Memory

Maithilee Kunda[1] MKUNDA@VANDERBILT.EDU
Julia Ting JULIA.TING@GATECH.EDU
School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30308 USA

Abstract

We present a new computational model that addresses resource limitations in working memory during visual search. In contrast to previous computational models of visual search, which assume unlimited and instantaneous access to a visual search template stored in working memory, our model provides a new mechanism for sequential, partial access to complex search templates. Using the Embedded Figures Test (EFT) as our task domain, we show that small variations in this mechanism for attention-based template access can have large effects on search performance. We also suggest a new computational explanation for how this mechanism might explain individual differences in "field independence," the cognitive construct that the EFT is intended to measure. Finally, we discuss the implications of our results for research in AI and cognitive science.

1. Introduction

Visual search is a fundamental process of intelligence, guiding how agents sample information from the visual environment, solve problems, and achieve goals. Understanding the information-processing mechanisms that underlie successful visual search is crucial for developing robust and efficient AI systems, ranging from sensor networks used for surveillance to mobile robots operating in complex environments.

In research on human cognition, search is studied in the context of visual attention. Attention is often classified according to whether it involves shifts that are (1) overt (externally observable) vs. covert (not externally observable) and (2) bottom up (stimulus driven) vs. top down (goal driven). Although artificial search tasks can be constructed to isolate these different processes, it is generally agreed that complex, real-world search tasks integrate all of them in some way.

A less widely studied aspect of visual search is how internal memory interacts with these kinds of transient attentional processes to produce observed search behaviors (Hutchinson & Turk-Browne, 2012). There are many roles that memory plays in visual search, including spatial memory of previous search locations/patterns (Peterson et al., 2001) and memory-based cuing and priming effects on the salience of different stimuli in the visual field (Desimone, 1996). In

[1] Current address: Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235 USA.

this paper, we focus on how variations in access to *the stored memory representations of target items* affect overall search performance.

The stored memory representations of target items can come in drastically different flavors, depending on the search task. Simple visual searches might use an iconic visual representation of the target, or *search template*, but one can imagine more complex visual search tasks in which the target is represented using a combination of visual, semantic, phonological, or other types of information, such as looking for "something to eat" in the wilderness or looking for "something that rhymes with cat" in a picture book. We narrow the scope of our investigation by focusing on the first case: search using an iconic visual template to represent the target.

In certain visual searches, the idea that working memory stores and uses an iconic visual template of the target is borne out by evidence that visual information is organized at the level of objects and not at the level of individual features (Luck & Vogel, 1997). Furthermore, individuals with high working memory capacity for visual information do better on visual search tasks than individuals with lower capacity (Reijnen, Hoffmann, & Wolfe, 2014).

Two computational models have examined the use of iconic visual templates in visual search (Rao, Zelinsky, Hayhoe, & Ballard, 2002; Zelinsky, 2008). In these models, both the template and the environment are represented as responses to a set of spatiochromatic filters. Correlations between the filter-response arrays of the template and of the environment are used to generate measures of visual salience that capture possible locations of the target in the environment. Both models propose specific, but different, mechanisms for using these salience maps to generate gaze shifts to desired locations. These models are both motivated by observations of human behavior in which, instead of generating a single eye movement towards the highest-salience location, several eye movements are directed in sequence to different locations in the search environment, often as a successively improving series of approximations to the final target location.

The first model proposes that several salience maps are computed at successively finer spatial scales, using filters that get smaller relative to the search environment. Gaze shifts are directed not to the most salient location in the final salience map, but rather to the most salient location in the current map (Rao et al., 2002). In other words, this model assumes that the time of computing each salience map is nontrivial when compared to the speed of executing gaze shifts, and so gaze shifts are generated as a gradually-improving sequence of approximate shifts towards the target location. The behavior of this model closely matches observations of saccades made by humans performing a simple search task to locate different objects on a tabletop.

The second model also proposes a series of approximate gaze shifts that move closer to the true target. However, these gaze shifts are directed to spatial averages of salient locations within a single salience map. This model accounts for the distracting effects of having multiple target-like elements in the environment, and it also includes mechanisms that inhibit the return of gaze to previously searched locations (Zelinsky, 2008).

Both models store a search template in working memory as an iconic visual representation. However, even though both models assume limited resources in accessing the search environment and salience maps, they assume that the entire template is available at every moment during the search process. In contrast, we explore what happens when the search template is *itself* subject to attentional deployments. We propose that this mechanism might explain some of the variability observed in human eye movements during search tasks.

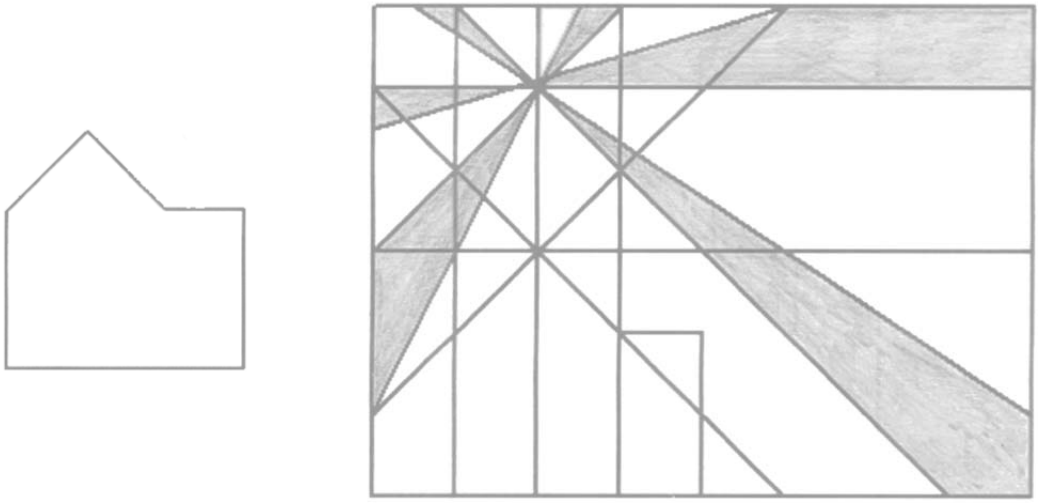

Figure 1. Example of problem similar to those found on the Embedded Figures Test. (Actual problems are not shown, in order to protect the security of the test.) The figure on the left is known as the "simple form." Solving an item requires locating the simple form somewhere inside the "test form" on the right. Rotating, scaling, or otherwise transforming the simple form to find it in the test form is not allowed. While taking the test (and unlike the example shown here), the simple form and test form are never simultaneously visible. The test taker *must* store a representation of the simple form in memory before attempting to search for it in the test form.

The basic question we address is how the mind deals with limited cognitive resources, which is a common theme across many different cognitive processes (Norman & Bobrow, 1975). For very simple search templates, assuming constant-time access to the full template seems reasonable. However, increasing the complexity of the search template is likely to have at least some effect on search performance.

We hypothesize that attention can be deployed internally to different spatial locations within visual working memory, to provide selective access to different parts of a stored search template. This hypothesis is analogous to how attention can be deployed covertly (i.e., without associated eye movements) to different locations of the visual field. Our hypothesis is also consistent with findings from mental imagery that visual working memory contains similar (and functionally useful) neural representations of visual information that are, in many ways, comparable with what is received during perception (Kosslyn et al., 1999; Kosslyn, Thompson, Kim, & Alpert, 1995; Slotnick, Thompson, & Kosslyn, 2005; Stokes, Thompson, Cusack, & Duncan, 2009).

The task domain that we study is the Embedded Figures Test (EFT), a widely used cognitive assessment in which participants must search for a simple geometric figure within a larger, more complex figure, as shown in Figure 1. We first describe the EFT in more detail, including the findings from human cognition that motivate our work. Next, we describe our computational model, the Attention-to-Template Visual Search (ATVS) model. Then we present experimental results from running different configurations of the model on the actual EFT. Finally, we close with a discussion of our results and their implications for understanding the computational mechanisms that contribute to visual search in intelligent systems.

2. The Embedded Figures Test

The Embedded Figures Test (EFT) was originally designed by Witkin (1950) as a measure of *field independence*, which refers to how well someone can differentiate an individual stimulus from background elements or patterns. Faster or more accurate searches on the EFT indicate greater field independence. Figure 1 shows an example of an EFT-like problem.

For each item in the original EFT, the examiner first presents the "test form," which is the complex figure to be searched (i.e., the search environment) and then presents a "simple form," which is the item to be found (i.e., the search target). Then the test form is once more presented to the subject before he or she begins the actual search. Witkin specifies that the simple form and test form should never be presented to the subject at the same time, but that the subject can ask to refer back to the simple form as needed (Witkin, Oltman, Raskin, & Karp, 1971). Interestingly, this procedure *requires* the subject to store the simple form in memory before searching for it in the test form. Performance is measured according to the time needed to complete each item.

There are several variants of the EFT currently in use (Ludwig & Lachnit, 2004). The most widely used are the Group Embedded Figures Test (GEFT) and the Children's Embedded Figures Test (CEFT). In order to be administered in a group setting, the GEFT uses a paper-and-pencil format, and subjects are given fixed time limits to complete three different sets of items (Witkin et al., 1971). The test is scored according to the number of items correctly completed within the time limits. Each page presents a complex form together with a letter indicating which simple form is to be found. The set of simple forms is printed on the back of the test booklet. This design enforces Witkin's specifications that (1) the test form is seen prior to the simple form for each item and (2) the test form and simple form are never simultaneously visible to the subject.

The CEFT was designed to be an easier, more engaging test than the original EFT for use with young children. The test introduces concrete shapes (e.g., houses, tents, strollers) for both the simple forms and test forms (Goodenough & Eagle, 1963). The CEFT is administered in a manner similar to the EFT, with a single examiner and a single subject (Witkin et al., 1971). Scores are recorded as the number of items correctly solved, although many research studies using the CEFT also record time to completion as a variable of interest.

In this paper, we use the GEFT for our computational experiments. However, our observations apply generally across all of the EFT variants, and so we use the abbreviation "EFT" to refer to this task domain as a whole throughout the remainder of the paper.

Studies of typically developing individuals have found that EFT performance is related to performance on other, similar disembedding-type tasks (Ghent, 1956). In addition, certain manipulations in administration formats, such as group vs. individual administration, differences in the coloration of test items, and memory requirements imposed by the task administration format can affect performance in significant ways (Jackson, Messick, & Myers, 1964). Interestingly, there have been substantial sex differences observed for EFT performance, although practice appears to reduce or remove these differences (Goldstein & Chance, 1965). Cultural differences in EFT performance have been observed as well (Kühnen et al., 2001).

Over the last few decades, many studies have found interesting patterns of differences in EFT performance between typically developing individuals and individuals diagnosed with autism spectrum disorder. These studies generally observe that individuals on the autism spectrum show superior EFT performance, in line with performance on other visual search tasks (Jarrold,

Gilchrist, & Bender, 2005), in the form of improved accuracy, shorter reaction times, or both (Jolliffe & Baron-Cohen, 1997; Shah & Frith, 1983). These differences appear to be related to differences in brain activity (Ring et al., 1999) and eye fixation durations (Keehn et al., 2009). Studies have also observed interactions between EFT performance and cultural differences in autistic populations (Koh & Milne, 2012).

The EFT and two other cognitive assessments—the Block Design test and the Raven's Progressive Matrices test—are often found to represent "peaks" of ability among individuals on the autism spectrum (Dawson, Soulières, Gernsbacher, & Mottron, 2007; Shah & Frith, 1993). All three of these tasks are visually presented and involve primarily visuospatial reasoning, as opposed to linguistic or semantic reasoning (Kunda & Goel, 2011; Kunda, McGreggor, & Goel, 2013). In the block design task, colored blocks must be put together to match a given pattern. In the Raven's task, a matrix of geometric figures must be completed with the correct missing figure. All three tests are widely used as cognitive assessments in clinical and scientific settings.

We believe that visual working memory plays an important role across all of these tasks. Understanding the specific mechanisms at work in each task will greatly improve the usefulness of these tasks as cognitive assessments, as well as our general understanding of visual cognitive processing. In previous work, we examined problem solving using visual representations on the Raven's test (Kunda et al., 2013). Here, we focus on the EFT to more closely examine the interplay between visual memory and visual search. Ultimately, we aim to develop integrated models that combine perception, memory, and reasoning across these and other tasks.

3. A Computational Model of Attention in Visual Search

Like previous computational accounts of visual search (Rao et al., 2002; Zelinsky, 2008), our model makes certain theoretical assumptions about the visual search process:

1. The search target is stored in working memory as an iconic visual template.
2. The search environment is also represented iconically as the pattern of perceptual activation in the visual field generated by the agent looking at the environment.
3. Visual salience is computed by comparing levels of correlation between 2D arrays representing the search template and different locations within the search environment.
4. The process of search involves directing eye movements towards a sequence of high-salience locations in the search environment until the target is found.

Previous models have assumed that the search template in working memory can be accessed in its entirety at any time. In contrast, the distinguishing theoretical commitment we make is that:

5. Due to limited cognitive resources, the template in working memory can be accessed only in part at any given time. Thus, the salience map available for informing eye movements uses only one small part of the template at a time.

This is consistent with Witkin's (1950) original EFT paper, which reports that individuals pick a "complex" part of the simple form to anchor their search at various points in the test form. They then try to trace the outline in the complex figure (Witkin, 1950). In other words, humans do not use the entire simple form at every moment during the search process. This observation about human behavior motivates the design of our Attention-to-Template Visual Search (ATVS) model.

3.1 Representations in the ATVS Model

The ATVS model takes as input PNG image files of the simple and test forms for each EFT problem, scanned directly from a paper copy of the test booklet. The system reads each input as a grayscale image and represents it as a two-dimensional array of binary true/false pixels. We use a manually chosen, image-specific grayscale value as a threshold to convert pixels from grayscale to binary values. Thus, both the search template and the search environment are represented as 2D black and white pixel arrays that capture image intensity.

Visual salience in the search environment is represented as a 2D array of similarity values, each denoting similarity between some part of the search template and some part of the search environment. Similarity is computed as the number of pixels in the intersection of the search template region and corresponding environment region divided by the total number of pixels in the search template, i.e., as the Jaccard coefficient.

The two models referenced earlier (Rao et al., 2002; Zelinsky, 2008) represent the search target and environment as arrays of spatiochromatic filter responses to image inputs, intended to mimic the filter-like responses of neurons in the human retina and early visual cortex. Both operate on real-world images. In our experiments, the EFT task domain presents relatively simple visual information (shaded line drawings on a blank background). Thus, we find that encoding each image as a black and white pixel array is sufficient. More generally, as long as the search template and environment are both represented as 2D arrays of the same type of information, then salience can be computed as an array of correlations between them. The rest of the model is agnostic towards the underlying representations used prior to calculating salience.

3.2 Processes in the ATVS Model

The ATVS model solves a single EFT problem at a time; no information is carried over from problem to problem. The system stores each search template as an ordered collection of non-overlapping visual *features*, each represented as a subimage of the original target image (line 1 in Table 1). For the current implementation, we defined these features manually. Features could be constructed automatically by assuming a fixed 2D subimage size or by defining boundaries according to a hierarchy of visual features (points, edges, lines, corners, etc.). Figure 2 (left) illustrates how the example simple form from Figure 1 is divided into features.

3.2.1 First-Stage Salience Calculation

The ATVS model uses these features to define a first-stage salience map according to all locations in the search environment where one of them is observed, i.e., locations that have high similarity values (lines 2-4 in Table 1). Figure 2 (right) shows an example of a first-stage salience map computed for the example problem shown in Figure 1, using the topmost corner of the simple form as the feature of interest. The system can search using any ordered collection of features that describe the search template, and it can also treat any given feature as the *anchor* that defines this initial, first-stage salience map.

Gaze is directed according to a random walk, without replacement, over these high-salience points in the search environment (line 5 in Table 1). At each location, the system computes

matches between the search template and locations in the search environment to determine whether a match occurs, which we call the second-stage matching process.

Table 1. Pseudocode for ATVS computational model, including both piecewise and comprehensive search mechanisms for use during the second-stage matching process.

solveItem(Image SimpleForm, Image TestForm, int anchor)

FIRST STAGE SALIENCE CALCULATION

1 Divide **SimpleForm** into features $\mathbf{f_1, f_2, \ldots, f_n}$, each represented as a subimage.
2 Choose $\mathbf{f_{anchor}}$ as the anchor feature for this search.
3 Compute a salience map representing the extent to which $\mathbf{f_{anchor}}$ matches different **(x,y)** locations in the **TestForm**, as,

$$\boldsymbol{salience}_{(x,y)} = \frac{\sum_{i,j}\left(\boldsymbol{TestForm}_{i+x,j+y} \cap \boldsymbol{f}_{anchor\,i,j}\right)}{\sum_{i,j}\left(\boldsymbol{f}_{anchor\,i,j}\right)}$$

4 For all **(x,y)** locations with $\mathbf{salience_{(x,y)}}$ greater than a threshold, add **(x, y)** to a list of high-salience points **HS**.
5 Randomly select and remove point $\mathbf{(x_{anchor}, y_{anchor})}$ from the list of points **HS**. (This is the point that *anchors* the second stage of the search.)

SECOND STAGE MATCHING PROCESS – PIECEWISE SEARCH

6 For each feature $\mathbf{f_i}$ in the list $\mathbf{f_1, f_2, \ldots, f_n}$, starting with $\mathbf{f_{anchor}}$:
7 Compute the expected position $\mathbf{(x_f, y_f)}$ of the feature $\mathbf{f_i}$ in the **TestForm**.
8 Search a small window around $\mathbf{(x_f, y_f)}$ for a match between $\mathbf{f_i}$ and the **TestForm**, using the **doComparison** function.
9 If **doComparison** is successful, go on to the next feature.
10 If there are no features remaining, the item is solved. The number of *fixations* equals the number of calls made to the **doComparison** function.
11 If **doComparison** is not successful, then select a new $\mathbf{(x_{anchor}, y_{anchor})}$ from the list of points **HS** (line 5). If **HS** is empty, the item fails to be solved.

SECOND STAGE MATCHING PROCESS – COMPREHENSIVE SEARCH

12 Search a small window around $\mathbf{(x_{anchor}, y_{anchor})}$ for the entire **SimpleForm** at once, using the **doComparison** function.
13 If **doComparison** is successful, the item is solved. The number of *fixations* equals the number of calls made to the **doComparison** function.
14 If **doComparison** is not successful, then select a new $\mathbf{(x_{anchor}, y_{anchor})}$ from the list of points **HS** (line 5). If **HS** is empty, the item fails to be solved.

doComparison(Image LocalTarget, Image TestForm, int (x_f, y_f), int window)

15 Calculate the maximum similarity between **LocalTarget** and **TestForm**, with **LocalTarget** positioned at location $\mathbf{(x_f, y_f)}$ in **TestForm**:

$$\boldsymbol{similarity} = \max_{\substack{x_f - window < x < x_f + window \\ y_f - window < y < y_f + window}} \frac{\sum_{i,j}\left(\boldsymbol{TestForm}_{i+x,j+y} \cap \boldsymbol{LocalTarget}_{i,j}\right)}{\sum_{i,j}\left(\boldsymbol{LocalTarget}_{i,j}\right)}$$

16 Return true if this **similarity** value exceeds a threshold and false otherwise.

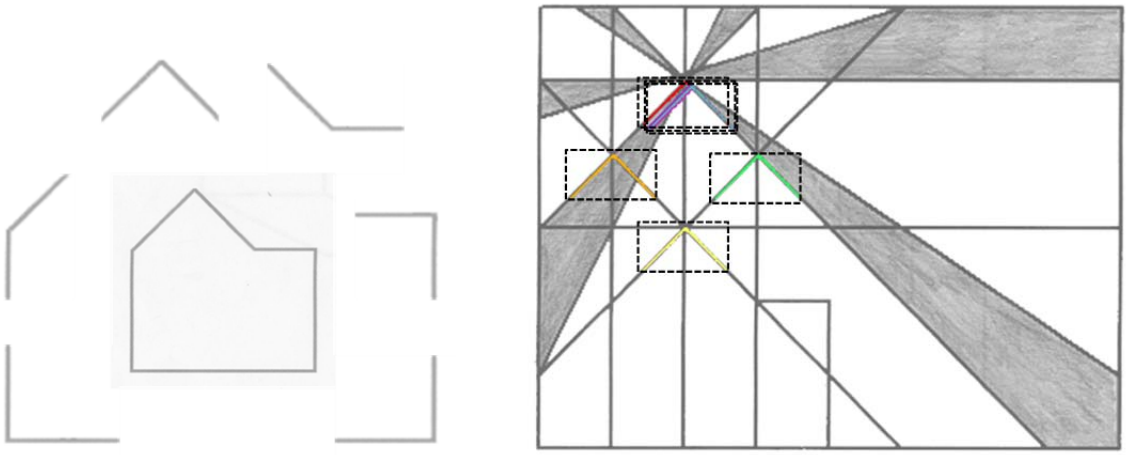

Figure 2. Left: Example simple form from Figure 1, carved into six visual features of interest with which to search test form. Right: Example of partial first-stage salience map created for problem shown in Figure 1. The anchor feature is defined as the topmost corner in the simple form, and the highest-similarity matching locations for this feature are shown in the test form. Dashed boxes indicate matches.

3.2.2 Second-Stage Matching Process

Whenever the ATVS model makes a comparison (lines 15-16 in Table 1), it simulates a fixation in the sense that the operation takes place for a short time period over a small, localized region of the visual environment and must be completed before it can move onto the next step. At each location, the system searches within a 20-pixel window of x-y alignments in order to determine whether a match exists, as determined by a thresholded similarity value. Once a match is found, the system ceases its search and goes onto the next feature or, if all features from the search template have been found, the search problem has been solved. We refer to this procedure as *Piecewise* search during the second-stage matching process (lines 6-11 in Table 1).

During Piecewise search, access to the search template is limited during both first-stage salience calculations and during second-stage matching. While this shows consistency across both search stages, it is difficult to isolate the contributions of access limitations to each individual stage. Therefore, we implemented an additional mechanism for the second-stage matching process, in which the model uses the entire search template at once (lines 12-14 in Table 1). We refer to this as *Comprehensive* search. In the Comprehensive variant of the ATVS model, access to the search template is limited only during first-stage salience calculations.

4. Experimental Evaluation

Using the ATVS model together with the EFT task domain, we aimed to test the empirical hypothesis that search performance depends, in part, on patterns of internally-directed attention to the search template stored in working memory. To be more precise, we hypothesized that the choice of anchor feature will have substantial effects on search duration, as measured by the number of fixations made by the model during the successful completion of an EFT problem.

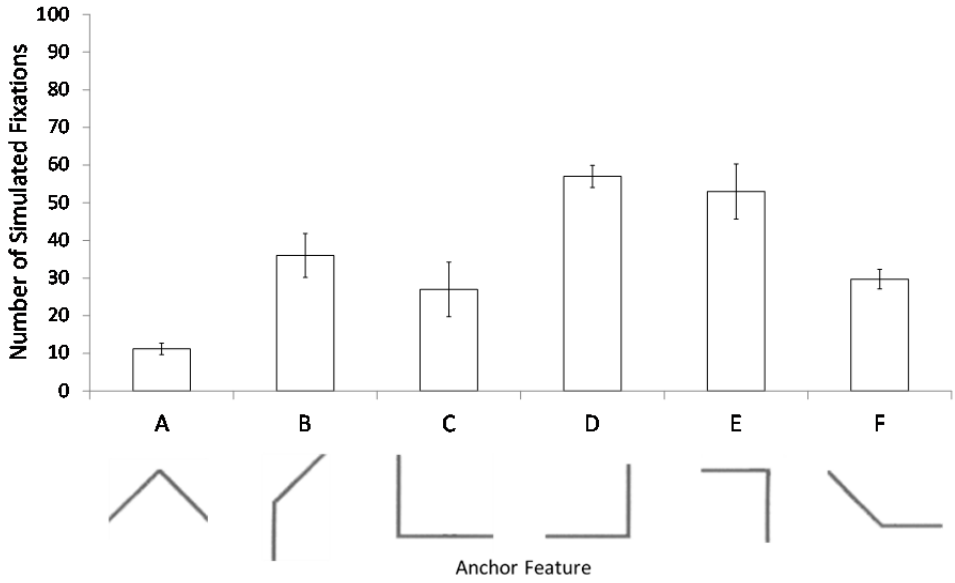

Figure 3. Results from the ATVS model for the example problem in Figure 1, using Piecewise search during the second stage matching process. This graph shows the mean number of fixations made across ten runs, with error bars indicating standard error, as a function of the choice of anchor feature.

The null hypothesis can be stated as: if variations in patterns of attention to the search template are not important, then we should expect to see only slight differences in search performance regardless of which feature is used as the anchor feature.

In this section, we present the results of two experiments. In Experiment 1, we tested the model using the default Piecewise matching strategy, in which access to the search template is limited during both stages of the search process. In other words, attention limitations are applied consistently to both the first stage salience calculations as well as the second stage matching process. However, this experiment alone cannot tell us anything about the relative contributions of attentional limitations during the first-stage salience calculations versus during the second-stage matching process. Thus, in Experiment 2, we tested the model using the Comprehensive matching strategy, in which access to the search template is limited only during the first-stage salience calculations. Experiment 2 provides a more conservative test of our hypothesis, and one that highlights the effects of attentional limitations during initial salience processing.

Because the ATVS model performs a random walk during part of its search process, outcomes are nondeterministic. For our experiments, we ran the model on each EFT problem ten times to obtain average measures of performance. We tested a few items using 100 runs, but results were not qualitatively different from results across ten runs. Therefore, we collected data across ten runs for each experiment.

4.1.1 Experiment 1: Effects of Limited Template Access with Piecewise Search

Figure 3 shows the mean and standard error for the number of fixations made by the model on the example problem from Figure 1, for various choices of anchor feature, using Piecewise search

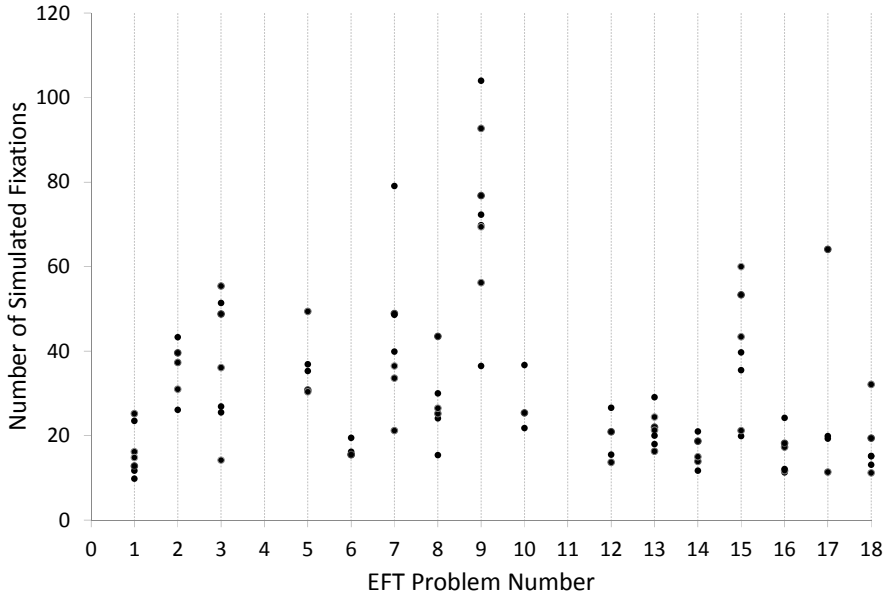

Figure 4. The number of simulated fixations made by the model for each problem from the EFT, using Piecewise search. Each data point indicates the mean over ten runs for each choice of anchor feature. Items 4 and 11 were not solved by the model.

during the second-stage matching process. Using the topmost point of the simple form as the anchor feature resulted in the lowest number of fixations, averaging about ten fixations to find the solution, while using the bottom right corner of the simple form resulted in the highest number of fixations, averaging about sixty fixations to find the solution. Results on this example problem provide positive evidence for our hypothesis that internally-directed patterns of attention to the template can have a considerable impact on search performance, potentially increasing the time taken to complete the search by a factor of six. However, these results were obtained from a single example problem authored by a member of our research team, and so we next present results from actual EFT problems.

Figure 4 shows results from running the ATVS model on EFT problems. Results are not shown for problems 4 and 11. Problem 4 was not solved by the model due to misalignments in the original drawings presented in the EFT booklet. We omitted problem 11 because it used a particular diamond-shaped simple form that the model had difficulty processing. In this graph, each data point represents the mean number of fixations for a given anchor feature, averaged across ten search trials. Because different problems on the EFT have different simple forms with varying numbers of features, the number of data points for each problem is not the same.

For some problems (e.g., 6, 13, and 14), the search times across different anchor features are clustered very closely together, suggesting that the choice of anchor feature has little impact on average search performance. However, for other problems (e.g., 3, 7, and 9), the choice of anchor feature caused a two- to five-fold difference in search performance. Thus, for many (though not all) problems on the EFT, results from Experiment 1 support our hypothesis that changing the pattern of attention to the search template has considerable effects on overall search performance.

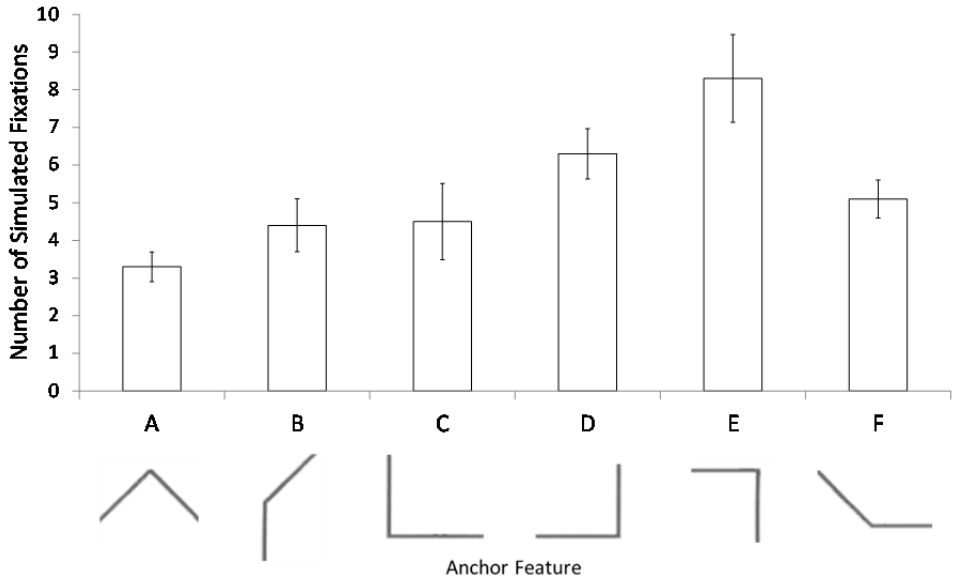

Figure 5. Results from the ATVS model for the example problem in Figure 1, using Comprehensive search during the second stage matching process. This graph shows the mean number of fixations made across ten runs, with error bars indicating standard error, as a function of the choice of anchor feature.

These results pinpoint a potential source of individual variation in search tasks that has not previously been studied: strategies for directing attention to different parts of a search template stored in working memory. The ATVS model does not implement concrete strategies for directing this kind of attention—*how* to choose the anchor feature—but it does provide upper and lower bounds on search performance as a function over each possible attentional deployment, i.e., every possible choice of an anchor feature.

Attention-to-template is an interesting source of individual variation in search tasks because it has nothing to do with more commonly cited factors such as spatial memory, perceptual speed, or general cognitive capacity. What results from Experiment 1 show is that, even with all of these perceptual and memory factors held strictly constant, differences in attention directed internally to a search template can produce very large differences in search performance. This finding seems to challenge Witkin's design of the EFT as a measure of field independence, but, as we discuss in Section 5, attention-to-template and field independence may be related.

4.1.2 Experiment 2: Effects of Limited Template Access with Comprehensive Search

Our second experiment assessed the performance impacts of attention to the template during the first-stage salience calculations. We used the Comprehensive search variant of the ATVS model for this experiment. Recall that Comprehensive search differs from Piecewise search by searching for the entire template at once during the second-stage matching process, instead of searching for it feature-by-feature as in Piecewise search. The total number of fixations produced by Comprehensive search will thus be smaller overall. The question we sought to answer in Experiment 2 is whether this decrease in overall fixations will be sufficient to eliminate the

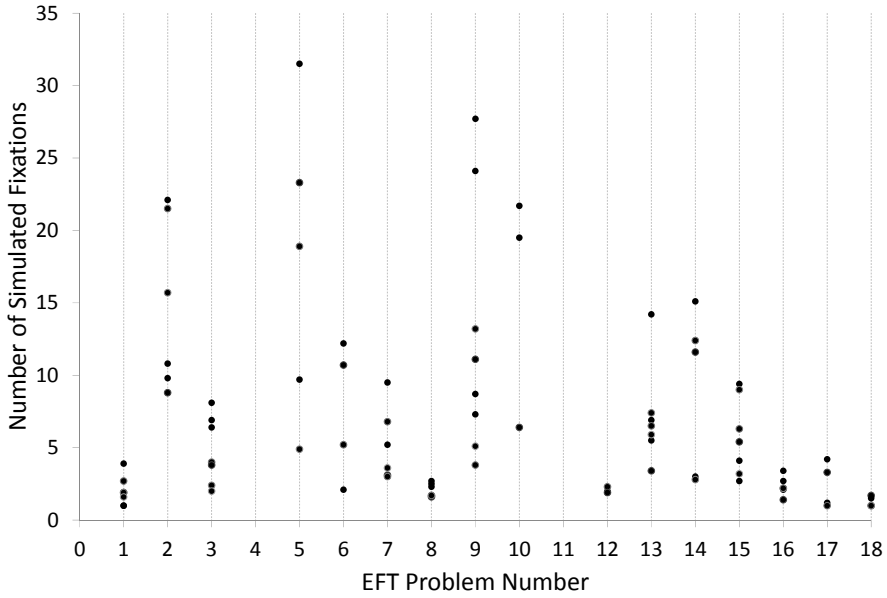

Figure 6. The number of simulated fixations made by the model for each problem from the EFT, using Comprehensive search during the second-stage matching process. Each data point indicates the mean over ten runs for each choice of anchor feature. Items 4 and 11 were not solved by the model.

effects of attention-to-template demonstrated in Experiment 1. In other words, if we limit the effects of attention to the first-stage salience calculations alone, will the choice of attentional deployments still have an effect on performance?

Figures 5 and 6 present the same results given for Experiment 1, except the ATVS model was using Comprehensive search instead of Piecewise search. As expected, the model generated far fewer fixations across all problems, down by nearly an order of magnitude. However, there were still considerable variations in performance as a function of the choice of an anchor feature. On the example problem, the anchor feature yielding the fewest fixations was again the topmost point of the simple form, averaging about three fixations for successful solution, as shown in Figure 5. The feature yielding the most fixations was the top right corner of the simple form, averaging about eight fixations. In Figure 6, problems 1, 8, and 16 from the EFT were among those showing the least effect of attention to the template. Problems 2, 5, and 9 all showed very large effects, with a three- to seven-fold difference between maximum and minimum values.

Like the results from Experiment 1, those from Experiment 2 also support our hypothesis that attention to the template can have a considerable effect on search performance. These results further illustrate that the effect is not due solely to the iterative nature of the second-stage matching processing using Piecewise search, but is also a consequence of attention to the template during the first-stage salience calculations. As both experiments present empirical results from our ATVS computational model, our findings are not directly interpretable in terms of human EFT performance. However, taken together with Witkin's (1950) early observations of people choosing a particular feature to "anchor" their search, our results provide a strong rational argument for more detailed human studies, as we discuss next.

5. Discussion

There are two main implications of our results. First, many computational theories and models of attention account for limited cognitive resources in perception and spatial memory but neglect resource limitations within visual working memory, particularly related to storage of and access to the search template. We have demonstrated a model that offers one solution to this resource limitation problem by providing access to different parts of the search template at different times. Our model successfully solves problems from a real test of visual search, the Embedded Figures Task, and also emulates observations that people internally choose an *anchor* with which to guide their initial search of the environment.

Second, we have shown that, even with the same search procedures and memory capacity, factors related to attention to the template can lead to considerable differences in overall search performance. In the EFT task domain, we have identified a potential source of individual variation not previously discussed in the literature: the choice of an anchor feature. Our results demonstrate that this choice can affect search results, by providing empirically derived upper and lower bounds on performance that result from different choices.

The immediate research question that ensues, then, is how the choice of an anchor feature is made by human subjects. More generally, how might intelligent agents with resource limitations in working memory go about choosing such features, during a given visual search task? The most basic answer is, perhaps, that selection is random. However, given the wealth of literature on top-down and bottom-up influences on attention and their adaptive role in many facets of intelligent experience, it seems likely that some combination of perceptual and cognitive factors drives the choice. Assuming, then, that the choice of anchor feature is not random but instead is somehow adaptive to task needs, we discuss two additional possibilities.

The first is that the anchor feature is chosen based on properties of the simple form alone, e.g., "Choose the most distinctive feature from the simple form to use as the anchor." However, two interesting observations come from Witkin's design of the EFT that make this unlikely. On the EFT, the simple form changes from one item to the next, and Witkin et al. (1971) found this to be an important aspect of the overall test administration. They observed that if the simple form was kept the same for several items in a row, then the test's discriminability lessened, as all participants would begin to show increased field independence, or the ability to easily find the simple form regardless of the test form's complexity.

Witkin et al. also specified that the test form should be presented first, and that the subject should spend some time inspecting it before looking at the simple form. On both the original EFT and the CEFT, the examiner asks the subject to describe the test form out loud to ensure that they are sufficiently attending to it: "During the initial 15-second exposure of each Complex Figure, the Subject should be asked to describe it in any way he pleases. The purpose of this procedure is to impress the organization of the Complex Figure upon the Subject" (Witkin et al., 1971, p. 17). This emphasis suggests that there is some kind of mental set induced by looking at the test form that is necessary to EFT items functioning in the intended way. If the anchor feature were determined purely by the simple form, then this early presentation of the test form would make no difference to the anchor feature selection aspect of the search strategy.

The second possibility is that the anchor feature, although representing a fragment of the simple form, is chosen based on properties of the *test form*. Interestingly, EFT performance using this

strategy would then vary inversely as a function of the complexity of the test form. If the test form is simple and contains little redundancy in visual information, then search will be faster. If, on the other hand, the test form is complex and contains many features that overlap with features in the simple form, then search will be slower. Choosing an anchor feature based on the test form instead of the simple form is somewhat counterintuitive; if one is searching for a target, it seems sensible to pick the most distinctive part of the target to use as the anchor. However, under the ATVS model, the optimal strategy is to choose the part of the target that is most distinctive in the search environment, even if it occurs multiple times in the target.

So what, then, is happening at a cognitive level during the EFT while a human subject gazes for 15 seconds (a *very* long time) at the test form before looking at the simple form? In line with studies of visual priming, we conjecture that the subject might be primed to attend to features of the simple form that are most frequent in the test form. This priming would actually result in the *worst* possible choice of anchor feature, not the best. Subjects who exhibit the least amount of this kind of perceptual priming would thus gain an advantage on the task. This is consistent with the EFT being a test of *field independence* if we define that term as increased freedom from this perceptual priming effect. This is also consistent with Witkin's observations that, if the same simple form is used over and over, the EFT's ability to discriminate field independence decreases.

6. Contributions

We have presented a computational model of visual search on the Embedded Figures Test (EFT) that exhibits a novel mechanism for dealing with limited cognitive resources, namely the use of sequential, partial access to a complex search template stored in visual working memory. We argue that this Attention-to-Template Visual Search (ATVS) model better represents resource limitations in visual search than do models that assume unlimited and instantaneous access to the search template at any time during the search process. Because the ATVS model was designed specifically to investigate this single aspect of visual search, it does not realistically model many other processes, such as:

- Representing and computing salience (Itti, Koch, & Niebur, 1998);
- Storing and incorporating information about saccade history (Zaharescu, Rothenstein, & Tsotsos, 2005);
- Biased competition among spatial, featural, and object information during calculations of salience (Lanyon & Denham, 2004);
- Integrating top-down and bottom-up attentional mechanisms (Aziz & Mertsching, 2008);
- Planning saccades to maximize information gain (Pomplun, Reingold, & Shen, 2003); and
- Using two stages of processing to initially guide and then refine search (Wolfe, 1994).

A complete account of intelligent visual search should incorporate all of these mechanisms. To this end, one goal of our research is to expand the capabilities of the ATVS model and integrate it with other computational models that address reasoning, goal maintenance, and other aspects of visual cognition (Kunda, 2015; Kunda et al., 2013).

The resulting architecture will be able to address problems from several different task domains, such as the EFT, Raven's Progressive Matrices, and the Block Design test. These three cognitive assessments together give an insightful picture of visual cognition in general, and they also

highlight some of the unique cognitive patterns exhibited by individuals with autism and other atypical neurocognitive profiles (Kunda & Goel, 2011). Using this architecture, one important area of future work will be to investigate how changes in a small number of cognitive factors (like attention to a search template) cause systematic changes in behavior across a suite of cognitive tasks. Such experiments will help scientists and clinicians better interpret the results of standard cognitive assessments like the EFT.

With regard to the EFT specifically, we have shown that differences in attention to the search template can cause substantial differences in search performance. We presented a new and interesting conjecture that this kind of attentional strategy not only contributes to general individual variation but directly ties into the core EFT construct of field independence. We proposed that when a subject first looks at the test form and then afterwards looks at the simple form, she or he is primed to attend to the part of the simple form that is most prevalent in the test form. The magnitude of this effect will directly influence the quality of the anchor feature that each subject chooses; the larger the priming effect, the worse the resulting anchor feature and overall performance. To our knowledge, no other computational mechanism at this level of detail has been proposed as a possible explanation for the construct of field independence.

To further investigate this idea, we aim to conduct a series of studies with human subjects that compare their behavior against the predictions of the ATVS model. There are many possible approaches. One is to track the eye movements of participants to identify whether they first search for an anchor feature and then trace out the rest of the search target. Another approach is to construct artificial EFT stimuli that systematically vary the number of times each feature is present in the search target and/or search environment and then observe effects on performance. Through continued efforts in both computational modeling and human studies, we hope to improve scientific knowledge about visual search and other key cognitive processes. This line of research will advance our understanding of human cognitive functioning, especially in the context of atypical cognitive development, and it will also contribute to the study of new computational methods for visual search in resource-limited AI systems.

Acknowledgments

We thank Sy Rashid for valuable contributions to this study. This research was supported in part by funding from the Georgia Tech Undergraduate Research Opportunities Program and through the National Science Foundation Expedition Award No. 1029679.

References

Aziz, M. Z., & Mertsching, B. (2008). Visual search in static and dynamic scenes using fine-grain top-down visual attention. In A. Gasteratos, M. Vincze, & J. K. Tsotsos (Eds.), *Computer Vision Systems*, 3–12. Berlin: Springer.

Dawson, M., Soulières, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science*, *18*, 657–662.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences, 93,* 13494–13499.

Ghent, L. (1956). Perception of overlapping and embedded figures by children of different ages. *The American Journal of Psychology, 69*, 575–587.

Goldstein, A. G., & Chance, J. E. (1965). Effects of practice on sex-related differences in performance on embedded figures. *Psychonomic Science, 3*, 361–362.

Goodenough, D. R., & Eagle, C. J. (1963). A modification of the Embedded-Figures Test for use with young children. *The Journal of Genetic Psychology, 103*, 67–74.

Hutchinson, J. B., & Turk-Browne, N. B. (2012). Memory-guided attention: control from multiple memory systems. *Trends in Cognitive Sciences, 16*, 576–579.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Jackson, D. N., Messick, S., & Myers, C. T. (1964). Evaluation of group and individual forms of embedded-figures measures of field-independence. *Educational and Psychological Measurement, 24*, 177–191.

Jarrold, C., Gilchrist, I. D., & Bender, A. (2005). Embedded figures detection in autism and typical development: preliminary evidence of a double dissociation in relationships with visual search. *Developmental Science, 8*, 344–351.

Jolliffe, T., & Baron-Cohen, S. (1997). Are people with autism and Asperger syndrome faster than normal on the Embedded Figures Test? *Journal of Child Psychology and Psychiatry, 38*, 527–534.

Keehn, B., Brenner, L. A., Ramos, A. I., Lincoln, A. J., Marshall, S. P., & Müller, R.-A. (2009). Brief report: eye-movement patterns during an Embedded Figures Test in children with ASD. *Journal of Autism and Developmental Disorders, 39*, 383–387.

Koh, H. C., & Milne, E. (2012). Evidence for a cultural influence on field-independence in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 42*, 181–190.

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., Ganis, G., Sukel, K. E., & Alpert, N. M. (1999). The role of Area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science, 284*, 167–170.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature, 378*, 496–498.

Kühnen, U., Hannover, B., Roeder, U., Shah, A. A., Schubert, B., Upmeyer, A., & Zakaria, S. (2001). Cross-cultural variations in identifying embedded figures comparisons from the United States, Germany, Russia, and Malaysia. *Journal of Cross-Cultural Psychology, 32*, 366–372.

Kunda, M. (2015). Computational mental imagery, and visual mechanisms for maintaining a goal-subgoal hierarchy. *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*, Atlanta, GA.

Kunda, M., & Goel, A. K. (2011). Thinking in pictures as a cognitive account of autism. *Journal of Autism and Developmental Disorders, 41*, 1157–1177.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Ravens Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research, 22-23*, 47–66.

Lanyon, L. J., & Denham, S. L. (2004). A model of active visual search with object-based attention guiding scan paths. *Neural Networks, 17*, 873–897.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Ludwig, I., & Lachnit, H. (2004). Effects of practice and transfer in the detection of embedded figures. *Psychological Research*, *68*, 277–288.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*, 44–64.

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. *Psychological Science*, *12*, 287–292.

Pomplun, M., Reingold, E. M., & Shen, J. (2003). Area activation: a computational model of saccadic selectivity in visual search. *Cognitive Science*, *27*, 299–312.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*, 1447–1463.

Reijnen, E., Hoffmann, J., & Wolfe, J. (2014). The role of working memory capacity in visual search and search of visual short term memory. *Journal of Vision*, *14*, 1073–1073.

Ring, H. A., Baron-Cohen, S., Wheelwright, S., Williams, S. C. R., Brammer, M., Andrew, C., & Bullmore, E. T. (1999). Cerebral correlates of preserved cognitive skills in autism: a functional MRI study of Embedded Figures Task performance. *Brain*, *122*, 1305–1315.

Shah, A., & Frith, U. (1983). An islet of ability in autistic children: A research note. *Journal of Child Psychology and Psychiatry*, *24*, 613–620.

Shah, A., & Frith, U. (1993). Why do autistic individuals show superior performance on the block design task? *Journal of Child Psychology and Psychiatry*, *34*, 1351–1364.

Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2005). Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral Cortex*, *15*, 1570–1583.

Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *The Journal of Neuroscience*, *29*, 1565–1572.

Witkin, H. A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality*, *19*, 1-15.

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1971). *A manual for the Embedded Figures Tests*. Consulting Psychologists Press.

Wolfe, J. M. (1994). Guided search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202–238.

Zaharescu, A., Rothenstein, A. L., & Tsotsos, J. K. (2005). Towards a biologically plausible active visual search model. In L. Paletta, J. K. Tsotsos, E. Rome, & G. Humphreys (Eds.), *Attention and performance in computational vision*, 133–147. Berlin: Springer.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*, 787–835.