Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/5512--20-10,114

Strategic Perception through Attentional Tuning

Andrew Lovett

NCARAI Branch
Information Technology Division

September 29, 2020

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
29-09-2020	NRL Memorandum Report	04-01-2019 – 03-26-2020

4. TITLE AND SUBTITLE

Strategic Perception through Attentional Tuning

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
NISE

6. AUTHOR(S)

Andrew Lovett

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
N2T4

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/5512--20-10,114

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

10. SPONSOR / MONITOR'S ACRONYM(S)

NRL-NISE

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Karles Fellowship

14. ABSTRACT

Develop a novel, computational approach to visual processing that is motivated by our understanding of visual attention in humans. Attention can be tuned to prioritize the visual features or spatial regions that are most relevant for performing a task, to minimize unnecessary processing of task-irrelevant information.

15. SUBJECT TERMS

Computer vision
Visual attention
Cognitive modeling

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Andrew Lovett
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	Unclassified Unlimited	6	**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited			(202) 767-0233

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

STRATEGIC PERCEPTION THROUGH ATTENTIONAL TUNING

INTRODUCTION

Motivation

A central challenge in visual processing is distinguishing relevant from irrelevant information. There is a vast amount of information potentially available even in a simple image. For example, if an image contains five objects, then binary spatial relationships can be computed between twenty object pairs, or more if the objects can be broken down into their parts. To address this representation problem, perceptual systems should be designed to process only the information that is necessary for performing a task. We believe the most efficient, and the most flexible, approach is to model the human visual attention system. Human attention can be tuned dynamically so that task-relevant objects are more likely to be selected for visual processing. For example, given the task instructions "Find the house to the right of the large, red house," humans can direct their attention first to a large, red object in the image, and then to a second object to the right of it. The objects selected, and the order in which they are selected, determine the representation that is constructed, ensuring that the representation is closely tied to the task at hand.

Objective

This research aims to develop a computational model of visual attention that mimics early visual processing in humans. The model should support attentional tuning, in which certain spatial locations or visual features (such as color) are highlighted, such that objects in those locations or possessing those features have a greater probability of being selected and processed. Attentional tuning should be guided by two factors: (1) explicit task instructions, as when the model is told to look for an object with particular features at a particular location; (2) selection history, for example, if the model has selected several green objects in the recent past, then it is more likely to select green objects in the future.

Background: Psychological Theory

In conducting this research, I explored the theory that human attention is guided by two complementary mechanisms: selection and enhancement (Lovett, Bridewell, & Bello, 2019a). Selection picks out an item, such as an object in the visual field, for further processing. In contrast, enhancement increases the likelihood that certain items will be selected in the future. Enhancement comes in two forms: spatial and featural. Spatial enhancement increases the likelihood of objects at particular locations being selected, whereas featural enhancement increases the likelihood of objects with particular features (for example, a particular color) being selected.

Selection and enhancement are distinct but closely interconnected mechanisms. After an object is selected, its location and features are enhanced, such that objects at the same location or with the same features are likely to be selected in the future. For example, after one looks at a red object, other red objects are more likely to capture one's attention. At the same time, people can deliberately direct enhancement based on task instructions, for example scanning through the area to the right of the red house to look for another object.

Manuscript approved August 6, 2020.

APPROACH

Framework

This research was conducted within ARCADIA, a computation framework designed to explore the interactions between attention, perception, cognition, and action (Bridewell & Bello, 2016). Models built in ARCADIA consist of a set of components and one or more attentional strategies. On each cycle of processing, the components operate in parallel, processing information and producing some output. One of the output items is selected as the focus of attention, according to the model's current attentional strategy. This focus is then broadcast to all components on the following cycle, which influences further processing.

To take a simple example, suppose an image contains three shapes: a red circle, a blue square, and a green triangle. Also suppose that the current attentional strategy prioritizes red objects (for whatever reason). If a component produces output describing one of the shapes as red, that output will be selected as the focus of attention. As a result, other information about the object, such as its shape, will be prioritized. The color, shape, and any other information will be integrated into an object representation that will be remembered over time—essentially, the model will remember that there is a red circle at a particular location. In contrast, because the other shapes have never been the focus of attention, there will be no explicit memory representation describing them.

ARCADIA also supports stimulus-response links, which are rules indicating that when certain preconditions are met, an action can be taken. An action might have a physical effect in a (virtual) environment, for example pushing a button. Alternatively, an action might only have a cognitive effect, for example storing information about an object in memory, or retrieving information about that object from memory.

Selection & Enhancement

Recall the claim that human attention relies on selection and enhancement mechanisms. Implementing selection in ARCADIA is unnecessary, as the framework already relies on selection to guide processing. Thus, I have focused on implementing enhancement in ARCADIA.

Spatial enhancement is implemented via an enhancement map, which records the degree of enhancement (as a number) at each location in the image. Objects that overlap regions of higher enhancement are more likely to be selected in the future. Enhancement surrounds recently selected objects, increasing the likelihood that they or other nearby objects will be selected in the future. At the same time, stimulus-response links can trigger scanning actions, in which enhancement is projected along a trajectory, as in the example of scanning to the right of one object to look for another.

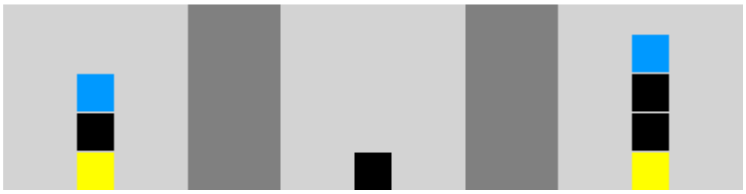
Featural enhancement is implemented in two ways. The simpler approach is to rely on a model's attention strategy. As described earlier, a strategy can prioritize a particular feature, such as a color—this means that if any component describes an object as having that color, the object is more likely to be selected as the focus of attention. In contrast, the more complex approach is to compute feature information across the visual field, checking each location's similarity to the currently enhanced features. This computation results in a second enhancement map, based on feature similarity, which can be added to the spatial enhancement map to determine the overall likelihood of objects being selected.

MODELS

This research has led to the development of two models. The first explores explicit task instructions—for example, being told to direct attention to the right of the large red object in the scene—whereas the second explores selection history, for example, attending to objects that are similar to those that have been recently selected. The models also vary in their scope. The first is a complete, but simplified model, in which the experimenters hand-crafted attentional strategies to match each set of task instructions. The second is a broader model, designed to operate as a visual front-end to task-specific models, but further work is needed to complete it.

Model 1: Visual Question-Answering

This model (Lovett et al., 2019) was designed to answer True/False questions about two-dimensional images. The model implements spatial enhancement via a component that performs relational sweeps. Given a referent object and a relation, such as "right of" or "within," the component projects an enhanced region that begins at the referent object and follows the appropriate trajectory. Objects that overlap the enhanced region and match the model's current featural priorities, as encoded in its attentional strategy, are selected as the focus of attention. As the model attends to objects, it encodes features of each object, as well as spatial relationships between pairs of sequentially selected objects. This information is stored propositionally in working memory and compared to a proposition representation of the True/False query to determine an answer.



True or False: There are two towers with a blue block at the top.

Fig. 1 — Typical question from the NLVR dataset (Suhr et al., 2017).

The model was evaluated on ten items from NLVR (Suhr et al., 2017), a dataset with simple images but conceptually challenging questions that require reasoning about presences, absences, spatial relationships, quantities, and commonalities or differences across image sequences—see Figure 1 for an example. For each item, the experimenters hand-crafted the appropriate attentional strategies and stimulus-response links. Overall, the model performed well, focusing on task-relevant objects in the images and answering all questions correctly. Because the attentional strategies were hand-crafted, running the model across the entire dataset was unfeasible. However, we are exploring a possible collaboration with a natural language understanding research group, who could support automatically constructing attentional strategies from English queries.

Model 2: General Visual Front-End

This model is designed to provide enhancement and selection capabilities across a variety of visual tasks. It includes an implementation of salience, a third attentional mechanism, which draws attention to object that stand out due to strongly contrasting with their surroundings (e.g., a single red object in a field of green objects). Salience, spatial enhancement, and featural enhancement maps are computed at every location in the visual field, following a gradated approach with fine-grained locational information in the center of the visual field and coarser information in the periphery—similar to the human visual system.

The three maps are summed to produce an overall activation level at each location in the visual field. Objects that overlap regions of high activation both (1) are more likely to be selected as the focus of attention, and (2) will tend to be selected more quickly.

We are in the process of testing the general visual front-end on two visual task models, each of which was previously developed with its own, task-specific front end: one for tracking moving targets, and a second for searching for a particular target in a field of distractors (Lovett, Bridewell, & Bello, 2019a, 2019b). By using the same front-end across these task models, we hope to demonstrate its generality.

CONCLUSIONS

The ongoing work on the two models has demonstrated that it is possible to implement selection and enhancement in a computational model, and that these mechanisms can help to focus visual processing, not only on task-relevant objects, but also on task relevant relationships between objects. Future work will explore the generality of this approach and evaluate its utility: the degree to which focusing on task-relevant objects saves processing time.

Future work will also expand the scope of this project beyond the visual modality. A planned ONR-funded project for FY21 will integrate visual and auditory attention, so that selection in one modality can guide enhancement in the other. For example, after a sound is heard to the far right, spatial enhancement will be directed in that direction so that the corresponding visual object can be identified, after which the visual and auditory information can be integrated into a common object representation. By pursuing this project, which will require close collaboration with an auditory processing group at the Naval Submarine Medical Research Lab, we hope to demonstrate the broad applicability of this work across the Naval Research Enterprise.

REFERENCES

1. W. Bridewell, and P. Bello, 2016, “A Theory of Attention for Cognitive Systems,” Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems, Evanston, IL, 2016.

2. A. Lovett, W. Bridewell, and P. Bello, 2019a, “Selection Enables Enhancement: An Integrated Model of Object Tracking,” Journal of Vision, 19(23). doi:10.1167/19.14.23.

3. A. Lovett, W. Bridewell, and P. Bello, “Attentional Capture: Modeling Automatic Mechanisms and Top-Down Control,” Proceedings of 41st Annual Meeting of the Cognitive Science Society, Montreal, Canada, 2019b.

4. A. Lovett, G. Briggs, B. McClimens, W. Bridewell, and P. Bello, “Visual Question Answering Through Strategic Perception,” Presented at the Advances in Cognitive Systems Workshop on Cognitive Vision: Integrated Vision and AI for Embodied Perception and Interaction, Cambridge, MA, 2019.

5. A. Suhr, M. Lewis, J. Yeh, and Y. Artzi, “A Corpus of Natural Language for Visual Reasoning,” Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 217–223.