SANDIA REPORT
SAND2017-
Unlimited Release
Printed September 2017

MODELING HUMAN COMPREHENSION OF DATA VISUALIZATIONS

Laura E. Matzen, Michael J. Haass, Kristin M. Divis, Andrew T. Wilson

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods/

SAND2017-10543
Printed September 2017
Unlimited Release

MODELING HUMAN COMPREHENSION OF DATA VISUALIZATIONS

Laura E. Matzen, Michael J. Haass, Andrew T. Wilson, Kristin M. Divis
1461, 1463
Sandia National Laboratories
P. O. Box 5800
Albuquerque, New Mexico 87185-MS1327

Abstract

This project was inspired by two needs. The first is a need for tools to help scientists and engineers to design effective data visualizations for communicating information, whether to the user of a system, an analyst who must make decisions based on complex data, or in the context of a technical report or publication. Most scientists and engineers are not trained in visualization design, and they could benefit from simple metrics to assess how well their visualization's design conveys the intended message. In other words, will the most important information draw the viewer's attention?

The second is the need for cognition-based metrics for evaluating new types of visualizations created by researchers in the information visualization and visual analytics communities. Evaluating visualizations is difficult even for experts. However, all visualization methods and techniques are intended to exploit the properties of the human visual system to convey information efficiently to a viewer. Thus, developing evaluation methods that are rooted in the scientific knowledge of the human visual system could be a useful approach.

In this project, we conducted fundamental research on how humans make sense of abstract data visualizations, and how this process is influenced by their goals and prior experience. We then used that research to develop a new model, the Data Visualization Saliency Model, that can make accurate predictions about which features in an abstract visualization will draw a viewer's attention. The model is an evaluation tool that can address both of the needs described above, supporting both visualization research and Sandia mission needs.

ACKNOWLEDGMENTS

The authors would like to acknowledge our collaborators at the University of Illinois at Urbana-Champaign and Georgia Tech. Their work has added invaluably to this project. Our collaborators at UIUC include Deborah Cronin, who collected the eye tracking data discussed in this report, and Jim Crowell, who programmed the initial experiment. Initial faculty support was provided by Hank Kaczmarski and Camille Goudeseune, with Alejandro Lleras stepping into the role of UIUC PI later in the project. Our primary collaborators at Georgia Tech were student Meeshu Agnihotri and her supervisor, faculty member John Stasko. We also had many fruitful research discussions with faculty member Alex Endert and his students.

TABLE OF CONTENTS

1. Project overview 7

1.1. Outline of Technical Work 7

1.2. Summary 9

References 11

Appendix A: Using Eye Tracking Metrics and Visual Saliency Maps to Assess Image Utility ..13

Appendix B: Modeling human comprehension of data visualizations 22

Appendix C: A New Method for Categorizing Scanpaths from Eye Tracking Data…………….33

Appendix D: Patterns of Attention: How Data Visualizations are Read………………………...38

Appendix E: Influences of Task Goals on Viewing Abstract Data Visualizations……………...47

Appendix F: Data Visualization Saliency Model……………………………………………..67

EXECUTIVE SUMMARY

A key factor in designing effective algorithms and tools is presenting the data to the human user in a format that s/he can interpret and understand. Visualizations are a common way to present data to users because humans rely heavily on vision to navigate the world, and those same cognitive processes can be used to navigate through information space. However, as data sets and analyses become ever more complicated, presenting information to analysts in a way that they can comprehend becomes ever more challenging.

While there is a great deal of research devoted to data visualization methods and techniques, efforts to assess the effectiveness of the resulting visualizations for the end users remain rare. Prominent researchers have argued that "the creation of comprehensive models of human-computer cognitive processing should be a core component of the visual analytics effort, and is an essential prerequisite for success of visual analytics as a field" [1]. In this project, we addressed these needs by combining a bottom-up model of visual saliency (the Data Visualization Saliency, or DVS model) with top-down eye tracking studies of sensemaking in the context of abstract data visualizations. We supported the development of this model with a series of human subject experiments, tracking participants' eye movements while they interacted with various types of visualizations under different goals. This novel combination of evaluation and modeling techniques drawn from the cognitive science and information visualization literatures helps lay the scientific foundation for evaluating data visualizations from a human cognitive perspective. Better understanding both what attracts a user's attention and why places us on stronger footing for designing more effective visual representations. With data complexity far outstripping the power of our representations, this ability constitutes a strategic advantage as well as a deep theoretical contribution.

This project built on Sandia's unique combination of strengths in data science, cognitive science, and information visualization to address fundamental questions about comprehension of abstract data visualizations, while leveraging Academic Alliance funded collaborations with both the University of Illinois and Georgia Tech. These questions are critically important for advancing the field of visual analytics and for improving human performance in the numerous mission areas that rely upon visualizations to support analysis and decision making.

1. PROJECT OVERVIEW

Data visualizations are ubiquitous in research and national security domains, and professionals in a wide variety of fields rely on visualizations when making high-consequence decisions. However, very little is known about how to evaluate a visualization's effectiveness for an end user. As data sets become larger and more disparate, it is becoming increasingly difficult to develop effective techniques for displaying complex, abstract data. For several years, prominent researchers in the field of information visualization have been calling for models informed by cognitive science to aid in the design and evaluation of data visualizations and visual analytics tools [1].

Despite the clear need for methods for evaluating data visualizations that are grounded in human cognition, at the start of this project, few researchers had addressed this issue. The few studies of how users navigate through data visualizations focused on fairly simple metrics, such as the order in which people view the axes on a graph [2; see also 3, 4, 5]. One reason for the lack of progress in this area was that few institutions have collaborations between cognitive scientists and visualization researchers, and even fewer have such collaborations in addition to access to the subject matter experts and analysts who are the intended end users of many visualizations. Sandia is uniquely positioned to address this issue because of our strongly interdisciplinary teams, access to subject matter experts, and our need for better methods in this area, both for applications within Sandia and in projects for external customers. Sandia has a strong history of visualization research, data science, and research on human cognition and decision making. We have a growing portfolio of visual cognition research that we leveraged to address these crucial questions about how people navigate through visual information when reasoning about the data and drawing conclusions.

We aimed to address the gaps identified above by conducting a series of studies that helped to lay the scientific foundation for evaluating visualizations from a human cognitive perspective. The project had two main goals. The first was to develop models for assessing the bottom-up visual saliency of data visualizations, and the second was to conduct eye tracking studies to develop models of the top-down sensemaking strategies employed by users of data visualizations. We utilized a novel combination of evaluation and modeling techniques drawn from both the cognitive science and information visualization literatures. This research is advancing the state of the art for evaluating the utility of data visualizations and has had a broad impact both within and outside of Sandia, benefiting numerous other projects.

1.1. Outline of Technical Work

Human visual processing is guided by two parallel processes: bottom-up and top-down visual attention, also known as stimulus-driven and goal-oriented attention [6]. Bottom-up visual attention is captured automatically by the physical properties of a stimulus (e.g. contrast, color, motion) while top-down visual attention is allocated voluntarily and is driven by the viewer's goals and expectations (e.g. what information the person is looking for and past experience with where to find that information [7]). The cognitive processing underlying visual search is thought to have two main processes. In the first stage, which happens very rapidly when a person first sees an image, the visual cortex of the brain pre-attentively filters the stimulus, identifying the most visually salient regions (the regions with high bottom-up salience). The information obtained at this stage of processing is then used to guide top-down visual attention, in which the viewer processes information serially by moving his or her eyes from one region of interest to another [8]. Regions with high bottom-up saliency may or may not be relevant to the viewer's task and goals, so there is a constant interplay between the two neural systems that guide visual attention and eye movements [9]. By focusing on both of the cognitive processes that guide humans' interactions with the visual world, we aimed to advance the scientific theories of visual attention while also providing practical guidelines for visualization designers.

The neural processes underlying bottom-up and top-down visual attention are fairly well understood, but the vast majority of the work prior to this project focused exclusively on natural scenes, such as photographs [10]. There were existing bottom-up saliency models that can predict where a viewer will look in a photograph [11]. The first publication produced by this project drew on prior visual cognition research at Sandia [12, 13, 14] and showed how an existing saliency model, the Itti and Koch model [11] could be combined with eye tracking data to evaluate the utility of scene-like data visualizations (Matzen, Haass, Tran & McNamara [15], see Appendix A for full text).

Our next step was to apply the Itti model and other popular saliency models to abstract data visualizations, such as those that are commonly find in scientific reports, software and system user interfaces, and visual analytics tools. We found that the models that perform best for images of natural scenes tend to fail for abstract data visualizations (Haass, Wilson, Matzen & Divis [16], see Appendix B for full text). Through a detailed assessment of where and why the models failed for abstract visualizations, we began to develop the Data Visualization Saliency (DVS) model to enable more accurate predictions of where viewers will look in a visualization.

To support the development of the model, we conducted a series of eye tracking studies to assess how viewers navigate through abstract visualizations. Our goal was to incorporate new features into the model to account for the unique visual properties of abstract visualizations. These features needed to be realistic in terms of how the human visual cortex processes information (e.g., appropriate color maps), and they also needed to be structured so that the contents of visualizations could reliably be incorporated into the saliency model. A cross-validation approach in which the model's saliency predictions were compared to recorded eye movements was used to determine the utility of each feature. The validity of this framework was tested using existing metrics that have been developed for assessing the match between predicted patterns of eye movements and actual user eye movements [17]. An initial study developed a new method for using scanpath data to infer a viewer's high-level task (Haass, Matzen, Butler & Armenta [18], see Appendix C for full text). Our first study that focused specifically on top-down influences on viewing of data visualizations found that viewers disproportionately attend to text in visualizations (Matzen, Haass, Divis & Stites [19], see Appendix D for full text). Two subsequent studies focused on the influence of high-level tasks and prior experience on comprehension of visualizations. A manuscript describing these studies is in preparation. See Appendix E for the preliminary results.

The results of the eye tracking studies informed the development of the DVS model, which is the first model of its kind to draw on both top-down and bottom-up characteristics in relation to data visualizations. The model expands the dimensionality of existing bottom-up saliency models and generates accurate saliency maps that can be used for evaluating abstract visualizations. The final, published version of the model significantly out-performs existing saliency models when applied to data visualizations, typically by a standard deviation or more (Matzen, Haass, Divis, Wang, & Wilson [20], see Appendix F for full text). The model is available for download at: https://github.com/mjhaass/DataVisSaliency.git

1.2. Summary

The lack of evaluation methods informed by models of human cognition was a crucial gap in the science of data visualization, both in terms of scientific understanding and in terms of mission needs, that this project has taken major steps toward addressing. Although some research had previously addressed this problem for natural scenes and for scene-like visualizations, many mission-critical visualizations are based on abstract or multidimensional data that cannot be tied to a natural physical representation. These are more difficult to design and evaluate, and they are also more difficult for an end user to interpret. When interpreting a photograph or a scene-like visualization, a user can draw on a lifetime of experience with navigating the physical world. Until this project, there had been very little research on how users navigate through abstract information spaces. This area carries a substantially higher level of technical risk because of the diversity of representations and applications for abstract data visualizations, as well as the absence of the constraints imposed by natural scenes on humans' visual search and reasoning strategies.

We have made substantial process in addressing this gap by integrating information about how humans process abstract visualizations from the perspective of both bottom-up and top-down visual cognitive processing. The outcome of this line of work is a widely applicable tool that can be used by data scientists and visualization designers to assess the visual saliency of their data visualizations and to predict (and guide) the user's allocation of attention. This in turn will support the end users of these visualizations, providing them with better tools that will enable faster and more accurate reasoning and decision making.

REFERENCES

1. Green, T. M., Ribarsky, W., & Fisher, B. (2009). Building and applying a human cognition model for visual analytics. *Information Visualization, 8*, 1-13.
2. Goldberg, J. & Helfman, J. (2011). Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. *Information Visualization, 10*, 182-195.
3. Burch, M., Andrienko, G., Andrienko, N., Hoferlin, M., Raschke, M., & Weiskopf, D. (2013). Visual task solution strategies in tree diagrams. *Proceedings of the IEEE Pacific Visualization Symposium*, 169-176.
4. Huang, W. (2007). Using eye tracking to investigate graph layout effects. *Proceedings of the International Asia-Pacific Symposium on Visualization (APVIS '07),* 97-100.
5. Kim, S., Dong, Z., Xian, H., Upatising, B., & Yi, J. S. (2012). Does an eye tracking tell the truth about visualizations? Findings while investigating visualizations for decision making. *IEEE Transactions on Visualization & Computer Graphics, 18,* 2421-2430.
6. Pinto, Y., van der Leij, A. R., Sligte, I. G., Lamme, V. A. F., & Scholte, S. (2013). Bottom-up and top-down attention are independent. *Journal of Vision, 13*, 1-14.
7. Connor, C. E., Egeth, H. E., & Yantis, S. Visual attention: Bottom-up versus top-down. *Current Biology, 14*, R850-R852.
8. Wolfe, J. M. (2007). Guided Search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford.
9. Ogawa, T. and Komatsu, H. (2004). Target selection in area V4 during a multidimensional visual search task. Journal of Neuroscience. 24, 6371- 6382.
10. Gegenfurtner, A., Lehtinen, E., & Saljo, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*, 523-552.
11. Itti, L. & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*, 194-203.
12. Haass, M. J., Matzen, L. E., Stevens-Adams, S. M. & Roach, A. R. (2015). Methodology for knowledge elicitation in visual abductive reasoning tasks. To appear in *Lecture Notes in Artificial Intelligence.*
13. Haass, M. J., Matzen, L. E. & McNamara, L. A. Top-down saliency estimation for advanced imaging scenes using pixel statistics. Paper presented at the European Conference on Eye Movements, August 2015.
14. Matzen, L. E., Haass, M. J., McNamara, L. A., Stevens-Adams, S. M., McMichael, S. N. (2015). Effects of professional visual search experience on domain-general and domain-specific visual cognition. To appear in *Lecture Notes in Artificial Intelligence.*
15. Matzen, L. E., Haass, M. J., Tran, J., & McNamara, L. A. (2016). Using eye tracking metrics and visual saliency maps to assess image utility. *Electronic Imaging*, *2016* (16), 1-8.
16. Haass M.J., Wilson A.T., Matzen L.E., Divis K.M. (2016) Modeling Human Comprehension of Data Visualizations. In: Lackey S., Shumaker R. (eds) *Virtual, Augmented and Mixed Reality. VAMR 2016. Lecture Notes in Computer Science*, vol. 9740. Springer, Cham.
17. Borji, A., Tavakoli, H. R., Sihite, D. N., & Itti, L. (2013). Analysis of scores, datasets, and models in visual saliency prediction. *IEEE International Conference on Computer Vision*.
18. Haass, M. J., Matzen, L. E., Butler, K. M., & Armenta, M. (2016, March). A new method for categorizing scanpaths from eye tracking data. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 35-38). ACM.

19. Matzen, L. E., Haass, M. J., Divis, K. M., & Stites, M. C. (2017, July). Patterns of Attention: How Data Visualizations Are Read. In *International Conference on Augmented Cognition* (pp. 176-191). Springer, Cham.
20. Matzen, L. E., Haass, M. J., Divis, K. M., Wang, Z., & Wilson, A. T. (in press). Data Visualization Saliency Model: A Tool for Evaluating Abstract Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics*.

APPENDIX A: USING EYE TRACKING METRICS AND VISUAL SALIENCY MAPS TO ASSESS IMAGE UTILITY[1]

Abstract

In this study, eye tracking metrics and visual saliency maps were used to assess analysts' interactions with synthetic aperture radar (SAR) imagery. Participants with varying levels of experience with SAR imagery completed a target detection task while their eye movements and behavioral responses were recorded. The resulting gaze maps were compared with maps of bottom-up visual saliency and with maps of automatically detected image features. The results showed striking differences between professional SAR analysts and novices in terms of how their visual search patterns related to the visual saliency of features in the imagery. They also revealed patterns that reflect the utility of various features in the images for the professional analysts. These findings have implications for system design and for the design and use of automatic feature classification algorithms.

[1] Matzen, L. E., Haass, M. J., Tran, J., & McNamara, L. A. (2016). Using eye tracking metrics and visual saliency maps to assess image utility. *Electronic Imaging*, *2016* (16), 1-8.

Introduction

Human visual processing is guided by two parallel processes: bottom-up and top-down visual attention, also known as stimulus-driven and goal-oriented attention [1]. Bottom-up visual attention is captured automatically by the physical properties of a stimulus (e.g. contrast, color, motion) while top-down visual attention is allocated voluntarily and is driven by the viewer's goals and expectations (e.g. what information the person is looking for and past experience with where to find that information [2]). The cognitive processing underlying visual search is thought to have two main processes. In the first stage, which happens very rapidly when a person first sees an image, the visual cortex of the brain pre-attentively filters the stimulus, identifying the most visually salient regions (the regions with high bottom-up saliency). The information obtained at this stage of processing is then used to guide top-down visual attention, in which the viewer processes information serially by moving his or her eyes from one region of interest to another [3]. Regions with high bottom-up saliency may or may not be relevant to the viewer's task and goals, so there is a constant interplay between the two neural systems that guide visual attention and eye movements [4].

Since the brain is so highly attuned to processing visual information, most human-computer interfaces rely heavily on the capabilities of the human visual system. A great deal of effort is devoted to finding ways to visualize information so that humans can understand and make sense of it. This is particularly challenging when the information is multidimensional, such as in visualizations with a temporal component. Once a visualization has been developed, assessing its utility for a human analyst can prove to be even more challenging than developing the visualization itself. Ideally, a visualization should draw the viewer's attention to the information that is most useful to the viewer's task. In other words, there should be overlap between the features that are visually salient and those that are most important from a top-down, goal-oriented perspective.

In this paper, we describe a study in which we assessed the utility of images by comparing viewers' eye movements to maps of visual saliency and image features. The project focused on Synthetic Aperture Radar (SAR) and Coherent Change Detection (CCD) imagery. SAR is used in a variety of surveillance and mapping applications and the radar data is converted into a two-dimensional image (see Figure 1) for use by human analysts [5]. CCD images (Figure 2) are created by co-registering SAR images of the same scene and measuring changes in coherence that can reveal changes that have taken place in the scene over time [6].

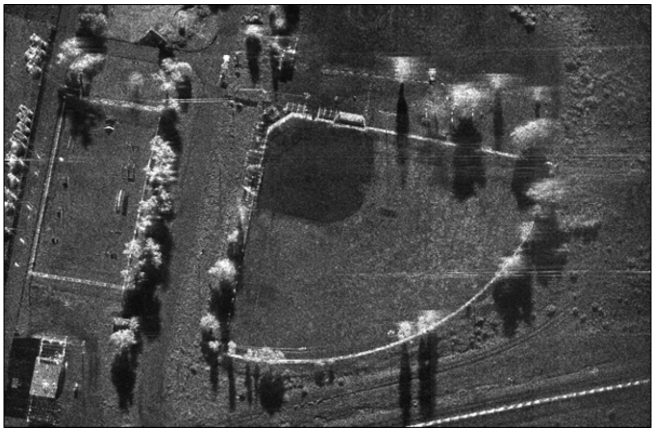


Figure 1. Synthetic Aperture Radar (SAR) image of a baseball diamond. Image courtesy of Sandia National Laboratories, Airborne ISR.

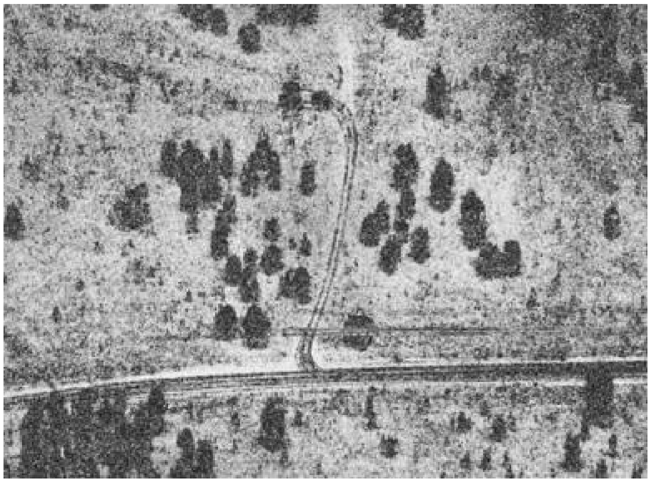


Figure 2. Coherent Change Detection (CCD) image highlighting several changes between images taken of the same scene at two different times. Image courtesy of Sandia National Laboratories, Airborne ISR.

Applied Studies of Imagery Analytic Workflows

The work described in this paper is part of an interdisciplinary family of research activities, in which Sandia National Laboratories researchers are examining how computational technologies influence the performance of professional imagery analysts. In this context, *imagery analysis* describes the perceptual and cognitive work of evaluating features of interest captured in two-dimensional images generated from remotely sensed data.

Visual inspection of imagery is an important component of work in a wide range of domains, from medical diagnostics to tactical military planning. However, the technologies used in imagery analysis have changed dramatically over the past couple of decades. Even as recently as the 1990s, "hardcopy" imagery and light tables comprised the major tools of imagery analysts. Importantly, the standards that express nominal thresholds for the detectability of feature classes in image products are rooted in psychophysical studies with imagery analysts using the hardcopy tool suite [7].

These days, however, computational or "softcopy" platforms are the main tools of imagery analysis. In many government workplaces, for example, light tables have disappeared as organizations have wholeheartedly embraced desktop computing systems and imagery analytic software. In a complementary fashion, computers have facilitated the development of image processing algorithms that can highlight or emphasize different features in a scene; for example, by exploiting changes in waveform characteristics to reveal ground changes in a scene-something that CCD imagery does very well. In short, the entire technological suite of imagery analysis has evolved dramatically over the past twenty years, with a wide array of electronic platforms and new image products available to support analytic workflows.

The imagery analytic revolution has raised questions about the functional equivalence of hardcopy vs. softcopy imagery for human visual detection tasks. A related issue is assessing the degree to which emerging image products might be used to support particular analytic workflows or feature detection goals. Finally, the rapid evolution of softcopy imagery also creates opportunities to examine how people interact with various types of image products as they are performing the visual cognitive work of professional imagery analysis. Of particular importance is the acquisition of perceptual skills, as people learn to "read" different types of imagery. We are particularly interested in understanding how imagery analysts learn to focus on the most valuable regions of an image product in relation to top-down analytic goals; and how these top down goals interact with bottom-up sensory and perceptual events driven by qualities of a given image product. Understanding these micro-processes is critical if we are to understand how people interact with imagery to establish a plausible narrative about the meaning of events captured in an image - for example, the import of footprints and tire tracks indicative of human activity in a rural area.

Current Research

The objective of this project was to identify which features in SAR and CCD imagery drew the attention of experienced and novice analysts during a visual search and decision making task. Our aim was to inform system design by identifying differences in search patterns between groups with varying levels of experience and relating those patterns to features in the imagery and their visual saliency.

SAR imagery is well-suited for this type of investigation for several reasons. First, SAR and CCD images are superficially similar to optical imagery, but extensive training is required for analysts to learn to interpret SAR phenomenology correctly. This creates unique advantages for studying the influence of experience and top-down visual attention on visual search behavior. Professional imagery analysts who work with SAR perform visual search tasks using SAR and CCD images on a daily basis, developing extensive expertise and efficient visual search and decision making strategies. At the same time, there are many true novices who have never seen SAR or CCD images, yet the similarity between SAR imagery and optical images enables novices to complete visual search tasks despite their lack of domain-specific experience. Second, several feature detection algorithms have been developed for SAR and CCD images. These algorithms can identify specific terrain features and image regions that are particularly useful (or not useful) to the visual search task. This allows us to map the participants' gaze patterns against image features with high or low importance from the perspective of top-down attention. Finally, prior research has shown that visual saliency maps designed for optical imagery, such as the tool developed by Itti and Koch [8], are also applicable to SAR and CCD images because of their scene-like properties [9]. This allows us to contrast the participants' gaze maps with maps of the bottom-up visual saliency of the images. All of these characteristics make SAR a particularly useful domain for studying differences in visual search between experienced and inexperienced viewers, and how those differences relate to properties of the images.

In the study, we collected behavioral and eye tracking data from three groups of participants with varying levels of experience with SAR imagery, ranging from true novices to professional SAR imagery analysts. The participants completed a visual search and decision making task in which they were asked to search SAR and CCD images for targets. The targets were specific types of changes within the scenes. The gaze maps collected from the three groups of participants were then contrasted with visual saliency maps and with maps of automatically segmented terrain features. We also conducted an exploratory analysis in which the gaze maps were compared to a metric of change susceptibility within the scenes, described in more detail below.

We hypothesized that in situations where the decision-relevant information was not the most visually salient information, novice viewers would be more likely to get distracted. In contrast, experienced analysts are likely to have developed strategies to discount salient but irrelevant visual features. We predicted that the experienced analysts would focus on the most task-relevant regions of the images, regardless of their visual saliency. Comparing the performance and eye movements of groups with varying levels of experience allowed us to investigate the influence of top-down visual attention on task performance and to explore the interplay between expertise and image utility.

Eye Tracking Study

Method

Participants

Twenty-four participants completed a target detection task using SAR images while their eye movements were recorded at 60 Hz using the FaceLab 5 Standard system and EyeWorks software. Eight of the participants were professional SAR analysts who conduct visual search tasks using SAR imagery on a daily basis. Eight were non-analysts who work with SAR images regularly, typically on a weekly basis. They had extensive knowledge of the domain, but do not typically engage in visual search tasks using the imagery. Most of the participants in this group were radar engineers who design and test SAR systems. We refer to this group as the "experienced non-analysts." The remaining eight participants were novices with no prior exposure to SAR imagery. All participants gave their written informed consent before participating in the study.

Materials

Participants completed a target detection task using 20 pairs of images. Each pair consisted of a SAR image and a CCD image of the same scene. The CCD image was created by co-registering SAR images of the same scene over time and measuring changes in coherence that can reveal temporal changes [6]. Essentially, the SAR image provided viewers with contextual information about the scene and the CCD image provided viewers with information about the presence or absence of targets in the scene.

Half of the 20 image pairs contained a target and half did not. The targets were the same types of targets that the professional SAR analysts look for in their daily work. The experienced non-analysts were also familiar with the nature of the targets and view them frequently, although not in the context of a visual search task. The novices were not familiar with the domain, so they were shown examples of targets before beginning the experiment. They received instructions about what to look for to determine whether or not a target was present in the scene.

Procedure

The participants completed a battery of general cognitive and visual search tasks in addition to the target detection task using SAR imagery [10]. In the target detection task, they were asked to stare at a fixation cross in the center of the computer screen. The cross remained on the screen for one second, and then one of the image pairs appeared on the screen. The SAR image was shown to the left of the fixation cross and the CCD image of the same scene was shown to the right of the fixation cross.

Participants were instructed to search the images for targets and to use a 1-4 scale to record their assessment of whether or not each scene contained a target. A response of "1" indicated that they were sure that there was not a target in the scene. A response of "2" indicated that they thought there was no target, but they were unsure. A response of "3" indicated that they thought there was a target present, but were unsure. A response of "4" indicated that they were sure that there was a target present. The SAR and CCD images remained on the screen until the participants responded or until 45 seconds had elapsed. The participants did not receive feedback about their answers until after the experiment was completed.

Results

Behavioral Results

The behavioral results showed that the professional imagery analysts were able to detect the targets more accurately than the novices and faster than both the novices and the experienced non-analysts. The analysts responded correctly to 74.4% of the trials, on average, with an average reaction time of 9.5 seconds. The experienced non-analysts responded correctly to 70.0% of the trials with an average reaction time of 14.5 seconds. The novice participants responded correctly to 56.9% of the trials with an average reaction time of 22.4 seconds.

One-way ANOVAs showed that the groups differed significantly in both their average accuracy ($F(2,21) = 4.62$, $p < 0.03$) and their average reaction times ($F(2,21) = 11.98$, $p < 0.001$). Post-hoc t-tests showed that the analysts had significantly higher accuracy ($t(14) = 2.95$, $p < 0.01$) and faster reaction times ($t(14) = 4.34$, $p < 0.001$) than the novices. The experienced non-analysts also had significantly higher accuracy ($t(14) = 2.14$, $p < 0.03$) and reaction times ($t(14) = 2.57$, $p < 0.02$) than the novices. The accuracy of the analysts and experienced non-analysts did not differ significantly ($t(14) = 0.73$), but the analysts had significantly faster reaction times ($t(14) = 2.93$, $p < 0.01$).

Eye Tracking Results

Two participants, one from the novice group and one from the experienced group, were excluded from the eye tracking data analysis due to noisy data. A region of interest (ROI) was demarcated around each target that contained the target itself plus a buffer intended to represent a person's useful field of view (approximately 90 pixels on each side of the target).

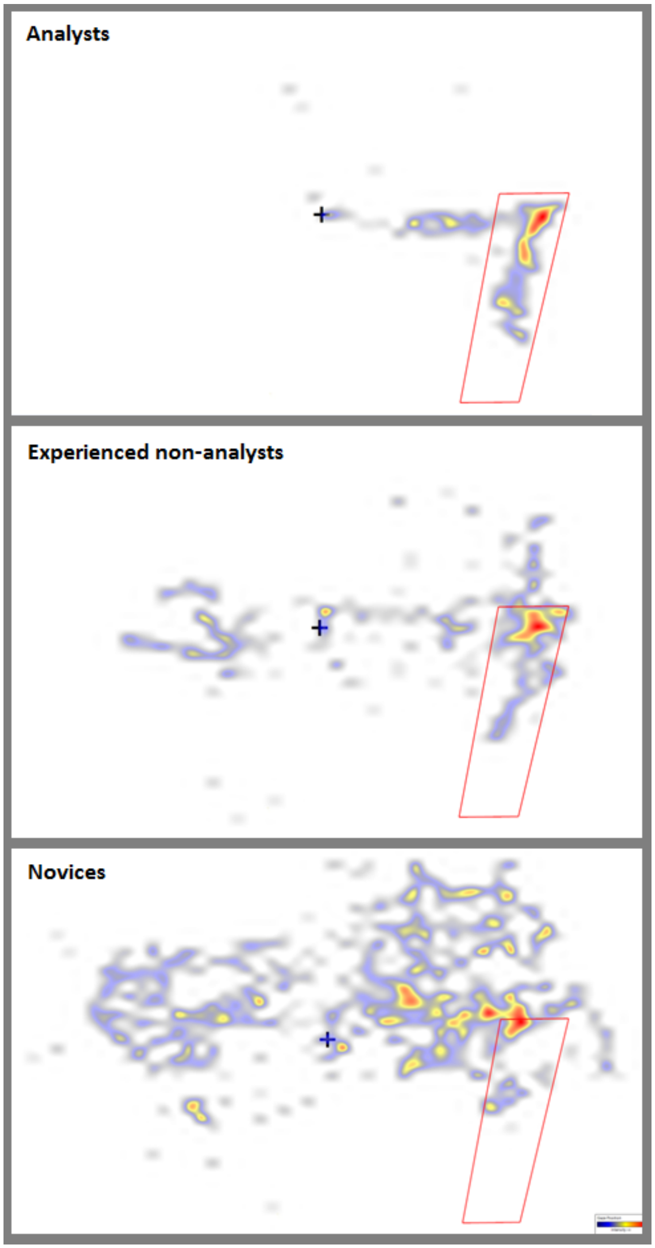


Figure 3. Gaze maps for each of the three groups of participants with the ROI indicated in red.

The time to first fixation in the ROI was calculated for each trial in which a target was present. The average time to the first fixation in the ROI was 5.3 seconds for novices, 3.0 seconds for experienced non-analysts, and 2.1 seconds for analysts. The difference between groups was significant ($F(2,19) = 9.21$, $p < 0.01$). Post-hoc t-tests showed that the experienced non-analysts and the analysts were both significantly faster than the novices ($t(12) = 2.41$, $p < 0.02$ and $t(13) = 4.36$, $p < 0.001$, respectively). However, the experienced non-analysts and the analysts did not differ significantly from one another ($t(13) = 1.53$, $p = 0.08$).

For each trial, we calculated the percentage of total fixations that occurred within the ROI. On average, 17.4% of the novice's fixations were in the ROI, compared to 25.3% for the experienced non-analysts and 38.9% for the analysts. The difference between groups was significant ($F(2, 19) = 8.08$, $p < 0.01$). Post-hoc t-test showed that the experienced non-analysts had a significantly higher percentage of fixations in the ROI than the novices ($t(12) = 2.47$, $p < 0.02$) and the analysts had a significantly higher percentage of fixations in the ROI than the experienced non-analysts ($t(13) = 2.13$, $p < 0.03$).

Discussion

Working within their domain of expertise, the SAR imagery analysts and experienced non-analysts were both more accurate in their responses than the novices, who had not viewed SAR imagery before taking part in the experiment. In addition to their high accuracy, the analysts were faster than experienced non-analysts and novices, both in terms of overall task reaction time and in terms of the time to first fixation in the ROI. The analysts were highly efficient in their ability to identify the ROI, typically fixating in the ROI within two seconds of stimulus onset. They devoted a higher proportion of fixations to the ROI than either of the other groups.

The efficiency of the analysts indicates that their visual search performance is driven by top-down visual processing. The analysts were able to rapidly triage the information in the imagery, zeroing in on the task-relevant information in the ROIs. In the analyses described below, we contrasted the gaze maps of the analysts and novices with other information about the content of the scenes, including bottom-up visual saliency and automatically detected terrain features. These analyses allowed us to further tease apart the contributions of bottom-up and top-down visual processing to the participants' visual search performance.

Comparison of Gaze Maps to Saliency Maps

In order to compare the visual search patterns of the participant groups to visual properties of the imagery, gaze maps were created for each stimulus using each group's tracking data. Following the approach of Wooding [11], the gaze maps were constructed by pooling the raw eye tracker samples over all subjects in each group (i.e. analysts, experienced non-analysts and novices) and accumulating a two dimensional Gaussian function at each point. The standard deviation of the Gaussian function was defined to equal a two degree field of view (90 pixels) at the average viewing distance.

Visual saliency maps for each stimulus where created using the Itti and Koch model [12] as implemented in Harel's Graph Based Visual Saliency Toolbox [13]. The Itti and Koch model decomposes images into three feature sets that are based on processes in the human visual cortex: color, orientation and intensity. These feature sets are constructed at multiple scales using Gaussian pyramids. Areas of the image with the greatest differences in features across scales are assigned larger saliency values while areas with smaller differences in features across scales are assigned lower saliency values. In this study, participants were viewing two images placed side by side on the screen. Because the two image products have different mean intensity levels, we calculated the saliency maps separately for each image product to avoid saliency artifacts at the image product boundary.

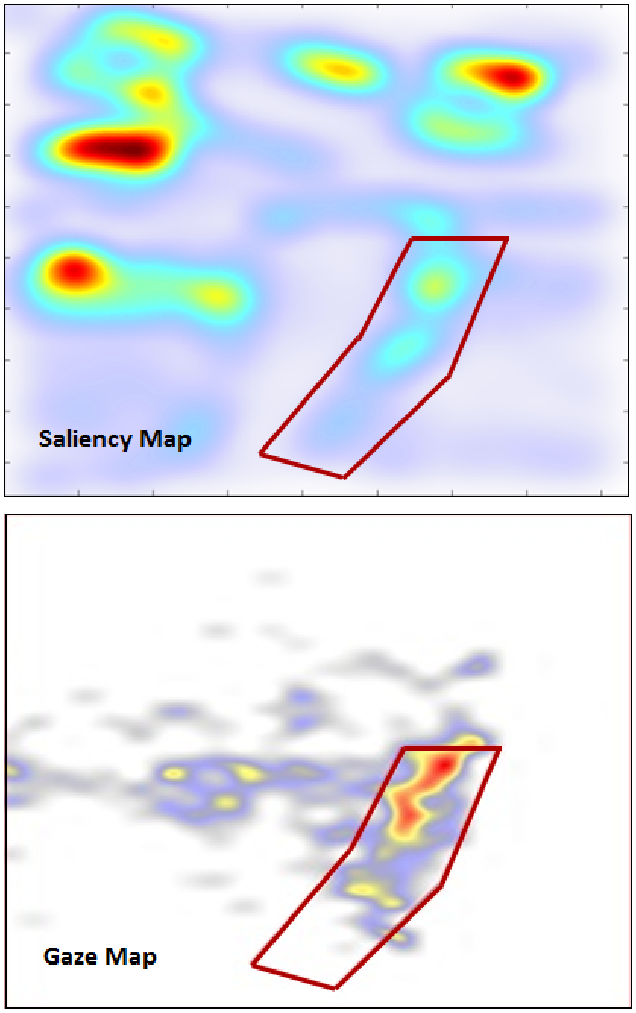


Figure 4. The top panel shows the saliency map for one of the CCD stimuli used in the study and the bottom panel shows the analysts' gaze map for the same stimulus. The ROI is indicated in red.

Results

For each of the 10 stimuli in the eye tracking study that contained a target, we calculated the percentage of the overall visual saliency that fell within the ROI around the target. Then, for each group of participants, we calculated the percentage of gaze observations that fell within the ROI for that stimulus. For all of the target-containing stimuli, an average of 17% of the total visual saliency fell within the ROIs. For the professional analysts, an average of 57% of the gaze observations fell within the ROIs, consistent with the behavioral finding that the analysts were very efficient in identifying the ROIs. The experienced non-analysts and novices had lower percentages of gaze observations in the ROIs, with 42% for the experienced non-analysts and 27% for the novices.

Correlations were calculated between the percentage of visual saliency in the ROI and the percentage of gaze observations in the ROI for each stimulus within each group of participants. The results showed that the correlation was significant for the novices ($R^2 = 0.71$, $p < 0.01$) and for the experienced non-analysts ($R^2 = 0.52$, $p = 0.01$). However, for the professional analysts, there was not a significant correlation between the percentage of saliency in the ROIs and the percentage of gaze observations in the ROIs ($R^2 = 0.02$).

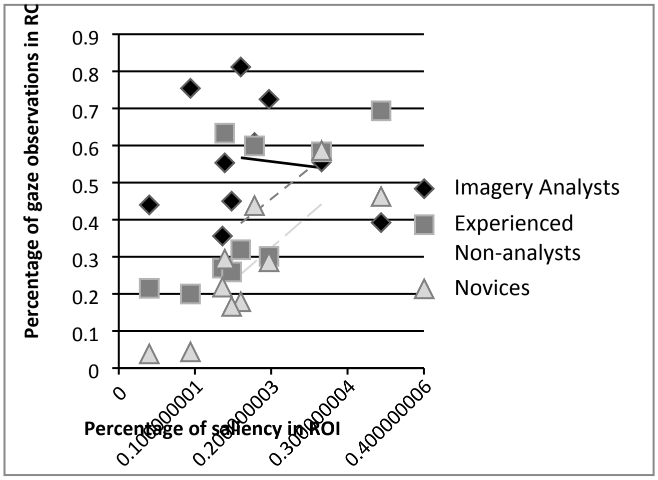


Figure 5.The percentage of gaze in the ROI versus the percentage of saliency in the ROI for each participant group for every stimulus that contained a target.

As discussed above, we hypothesized that professional analysts would rely on their past experience and on top-down visual attention to focus on the most task-relevant information, regardless of whether or not it was salient from a bottom-up perspective. The results of the eye tracking study and our comparisons between the gaze maps and saliency maps supported this hypothesis. To further explore the relationships between terrain features, visual saliency, and visual search, we compared the participants' gaze maps to automatically generated maps of image features. We chose to investigate two specific types of terrain features: SAR shadows and regions categorized as supporting change detection through a method called Index for Surface Coherence (ISC). These analyses and the preliminary results are described in the sections below.

Comparison of Gaze Maps and Terrain Features

SAR imagery has unique properties that support a variety of methods for automatic feature detection. For example, specific terrain features can be detected and labeled by automated image processing algorithms such as superpixel segmentation and classification [14, 15]. Superpixel segmentation groups pixels by capturing image redundancy [16, 17]. A new method known as ISC extends this capability by identifying image regions in which the terrain features are more or less conducive to change detection [18].

We chose to focus our analyses on two types of automatically detected terrain features. First, we contrasted the gaze maps with maps of SAR shadows. The shadows in SAR images have relatively low importance in target detection tasks, but have high visual saliency. We predicted that experienced analysts would ignore shadow regions while novices would be more likely to be distracted by their high visual saliency. Second, in an exploratory analysis, we contrasted the gaze maps with ISC maps representing regions of the images that were most supportive of change detection. We predicted that the analysts would devote more attention to the regions that were most likely to support change detection, particularly since they were being asked to complete a target detection task in which the targets were changes to the scene. In contrast, we predicted that novices would not have the experience needed to determine which regions were most valuable to completing the task, making them less sensitive to this metric.

Modulating Saliency Maps Using Terrain Features

In order to test the analysts' and novices' ability to ignore the highly salient but low value shadows, we calculated the overlap between the participants' gaze maps and the saliency maps with and without the shadows. First, algorithms were used to segment [14] the stimuli used in the eye tracking study into superpixels and to classify [15] the shadow superpixels.

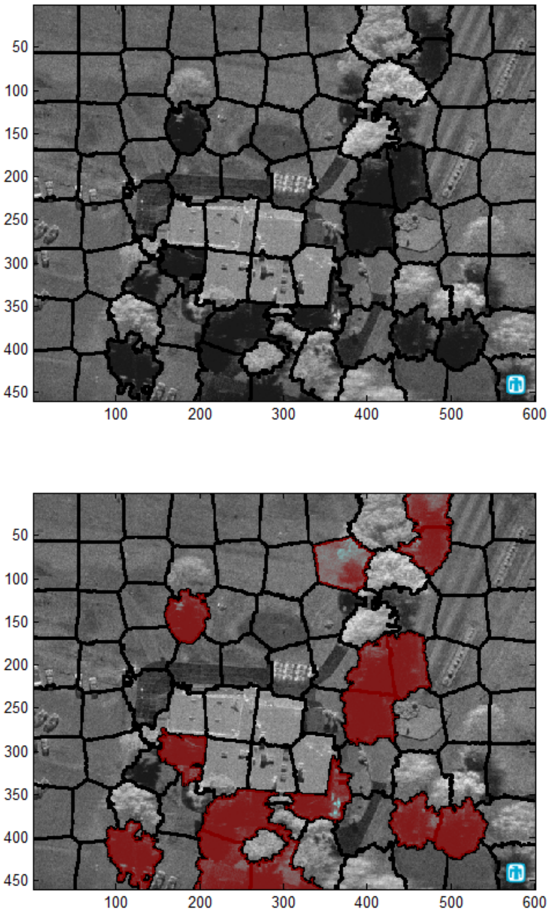


Figure 6. The top panel shows a superpixel segmentation of a scene and the bottom panel shows superpixels classified as shadow regions in red.

Next, modified saliency maps were created in which the superpixels identified as shadow regions were masked out, as shown in Figure 7.

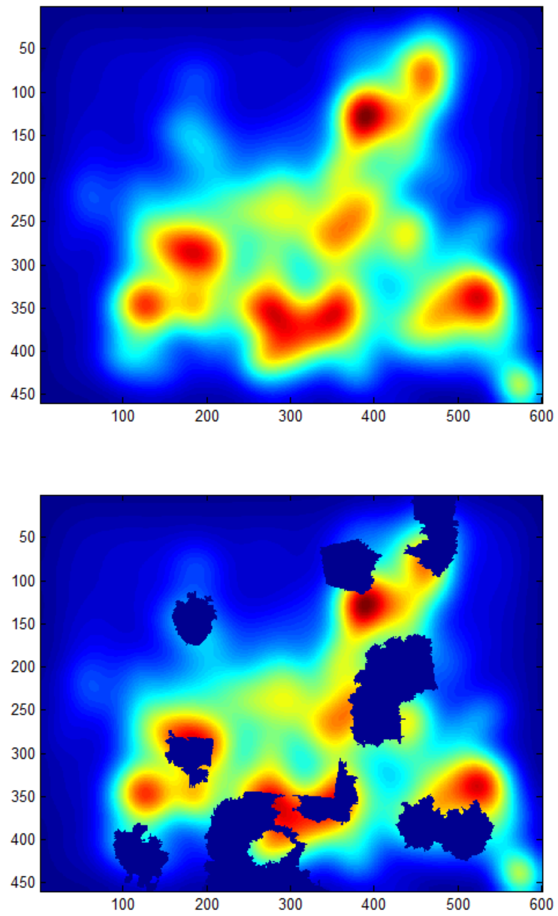


Figure 7. The top panel shows the visual saliency map created from the SAR image in Figure 6. The bottom panel shows the masking of the superpixels classified as shadow regions.

The gaze maps were compared to the original and masked saliency maps using the linear correlation coefficient (CC) metric. CC has been used in prior studies to measure performance of saliency estimation algorithms by comparing saliency maps to human gaze maps [19]. CC is a measure of the strength of a linear relationship between a gaze map (G) and a saliency map (S)

$$CC(G,S) = \frac{cov(G,S)}{\sigma_G \sigma_S} \quad . \qquad (1)$$

When CC is close to ±1, there is almost a perfectly linear relationship between the human gaze map and the predicted saliency map.

A subset of the eye tracking data (three analysts and three novices) was used to test the effects of masking shadows out of the saliency maps. For the analysts, masking the shadow regions improved CC agreement between saliency and gaze maps by a factor of 3.3 times. For the novices, masking the shadow regions *reduced* CC agreement by only 0.95 times.

These results provide further evidence to support our finding that professional analysts successfully relied on top-down visual attention, largely ignoring regions that were not relevant to the target detection task even if they were highly visually salient. The approach developed here could be applied for any other terrain features, allowing system designers to conduct a detailed analysis of how much experienced and novice users rely on each feature when completing a particular task. This could be a powerful method for assessing image quality by testing the relative contributions of each image feature to both the visual saliency of the scene and to the users' task performance.

Comparing Gaze Maps to the Index of Surface Coherence

As discussed above, CCD images provide a method for observing changes in a scene that would otherwise be undetectable to the human eye [20]. By using multiple SAR collects, the magnitude and phase difference between each collect can be utilized to detect changes in a SAR image. However, the method used to calculate this change product is agnostic to the underlying terrain on which the calculation is made. Some features (such as walls) are stationary and not susceptible to change, appearing as areas that cohere perfectly in the CCD images. Other features, such as vegetation, have low coherence due to their random geometries and continuously show up as changes in the CCD product. Both types of features can be distracting to an analyst or algorithm looking for changes of interest (i.e. areas of low coherence in the scene that typically have high coherence). Discerning changes of interest in natural scenes requires training for human analysts and a better understanding of the underlying terrain for algorithms.

A new method to address this issue creates maps of the Index of Surface Coherence (ISC) for SAR images. These maps can be used to mask a CCD product and eliminate the areas that do not support detection of changes of interest. To create these maps, a long-term observation of an area is utilized to acquire the underlying nature of the terrain. With many observations of the same area over a period of time, a stack of images can be created. By registering all of the images and taking the median of each pixel in the stack, a stable representation of the area is observed. Using a median radar cross section (RCS) and median CCD product, the terrain in the area can be classified according to its coherence properties. The median RCS (MRCS) and median CCDs (MCCD) images are segmented into superpixels using the SLIC superpixel segmentation, which allows a user to define how compact the superpixel appears and the number of superpixels in the image. This allows a user to create a nearly uniform grid of pixel groups [14, 17]. A truly uniform segmentation would provide pixel groups and reduce the computing complexity, but the pixels in those groups would be visually and statistically very dissimilar.

After the median MRCS and median MCCD images are segmented, a training process is used in which terrain types that support change detection are identified and a subset of superpixels capturing each terrain type is chosen. In this study, approximately 20 superpixels consisting of 500 pixels for each terrain type were selected. For each data type, a distribution curve is generated for both the MRCS and MCCD products. The distribution curve is generated by fitting common distribution types (Gamma, Beta, Log-Normal, Exponential, and Gaussian) to the each data type's scaled histogram data. The distribution type, distribution parameters, and scaling are saved to represent each terrain type.

With the training finished, new images can be evaluated by segmenting the image into superpixels and comparing each superpixel in the image to the previously trained data. For each superpixel in the image, its pixels are scaled and fit with the distribution according to each terrain types training data. The distribution curve of the superpixel is then compared to the terrain type's distribution curve using Kullback-Leibler (KL) Divergence to get a similarity score. Using probabilistic fusion [21, 22], the KL scores of the MRCS and MCCD images are translated into p-scores which can then be added despite the KL scores being statistically different. These added scores can then be used to form a heat map to indicate where an image is most likely to support change detection.

We conducted a proof-of-concept analysis in which an ISC map of one of the CCD images from the eye tracking study was compared to participants' gaze maps. To compare the image p-scores to the human gaze maps, we first created a set of 20 thresholded images (P) using the original p-score image and thresholding each pixel for thresholds 1,2,3,…20. We then calculated the CC metric for each thresholded image, P_i, compared to the gaze map from either the IAs or the novices.

$$CC(P_i,S_j) = \frac{cov(P_i,S_j)}{\sigma_{P_i}\sigma_{S_j}}$$

Where i = 1,2,…20; j = 1(analysts), 2(novices) (2)

At the lower thresholds, the maps show only regions that never change, while at higher thresholds the maps show regions with increasing susceptibility to change. This analysis showed that the CC metric peaked for novices at a p-score threshold of 2 while peaking for experts at a p-score threshold of 7. Although exploratory, these results indicate that the gaze maps of the novices were relatively insensitive to the likelihood that a particular region would support change detection. They devoted their attention to terrain features that did not provide much support for change detection and therefore had low p-scores in the ISC map. In contrast, the analysts devoted more attention to regions that had higher p-scores and were likely to support change detection.

Discussion

The results of this experiment revealed distinct differences between the visual search patterns of the participants in the three experience groups. Professional SAR imagery analysts were faster and more accurate in finding targets in a visual search task using SAR and CCD images. The results of the eye tracking study showed that the analysts were rapidly able to identify the ROI in the scenes containing targets and spent a significantly higher proportion of their time inspecting the ROI than the other groups of participants. The viewers with less experience, including non-analysts and true novices, spent more time viewing other regions of the images, which had a negative impact on their speed and accuracy.

To explore the relationships between the participants' gaze maps and the visual features of the imagery, we compared the gaze maps to bottom-up saliency maps and to maps of image features that were either irrelevant (shadows) or relevant (regions supporting change detection) to the task. While the gaze maps of the novices and experienced non-analysts were correlated with the bottom-up saliency of the images, the gaze maps of the professional analysts showed no such correlation. These results indicate that the less experienced groups were at least somewhat distracted by visual features that had high visual saliency but little relevance to the task. In contrast, the analysts focused their attention on task-relevant features, whether they were highly visually salient or not. In other words, the analysts' visual search processes appear to be driven primarily by top-down, goal-directed visual attention, while the less experienced participants were influenced more by bottom-up visual saliency.

The comparisons of the participants' gaze maps to automatically detected image features also supported this interpretation of the eye tracking data. We chose SAR shadows as an example of a visual feature that was highly salient but had little relevance to the task. When superpixels from shadow regions were masked out of the visual saliency maps, the match between the saliency maps and the analysts' gaze maps improved substantially. When the same masking was done for the novices, the match between the saliency maps and gaze maps was reduced. The comparison between the gaze maps and the ISC maps had a similar result. The highest match between the novices' gaze maps and the ISC maps was at a very low threshold, where the ISC map showed areas with little susceptibility to change. These areas are not very informative in a change detection task, but novice participants spent quite a bit of time looking at them. The analysts ignored those regions, focusing their attention on regions that were supportive of change detection and were therefore task-relevant.

The results of this study revealed information about what types of SAR and CCD image features are used by people with different levels of experience. By studying the professional analysts' approach to the visual search task and identifying the features and regions that they focus on, we were able to identify which features are most relevant to their real-world visual search tasks. This information can be used to inform system design and the design of new image products and image processing algorithms to support the analysts in their daily work. By comparing the professional analysts to experienced non-analysts and novices, we were also able to identify image features that might be distracting to less experienced viewers. This information can inform the training of new analysts. It can also help to validate new image processing algorithms. For example, the comparison between the participants' gaze maps and the ISC maps provided valuable feedback about the value of the ISC method for identifying regions that are relevant to the end users of the imagery. The threshold cutoffs identified by the gaze map comparisons can be used when deploying the algorithm to help analysts filter out potential false alarms.

The methods developed for this study could be applied in other domains to assess image quality in terms of how well the images support the end user's top-down goals. By approaching the problem from the perspective of human cognition, we were able to learn a great deal about the features of the images that did or did not support the end users' cognitive needs.

References

[1] Y. Pinto et al., "Bottom-up and top-down attention are independent," *Journal of Vision*, vol. 13, pp. 1-14, 2013.

[2] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology,* vol. 14, pp. R850-R852, 2004.

[3] J. M. Wolfe, "Guided Search 4.0: Current progress with a model of visual search," in *Integrated Models of Cognitive Systems,* W. Gray, Ed. New York: Oxford, 2007, pp. 99-119.

[4] T. Ogawa and H. Komatsu, "Target selection in area V4 during a

multidimensional visual search task," *Journal of Neuroscience*, vol. 24, pp. 6371- 6382, 2004.

[5] A. W. Doerry and F. M. Dickey, "Synthetic Aperture Radar," *Optics and Photonics News*, vol 15, issue 11, pp 28-33, 2004.

[6] A. W. Doerry, "SAR data collection and processing requirements for high quality coherent change detection," in *Proceedings of SPIE: Radar Sensor Technology XII, April 2008, Orlando, FL,* K. I. Ranney and A. W. Doerry, Eds. Society of Photo-Optical Instrumentation Engineers, 2008.

[7] J.M. Irvine, "National Imagery Interpretability and Rating Scales (NIIRS): Overview and Methodology," *Proc. SPIE 3128, Airborne Reconnaissance XXI*, 93, November 21, 1997

[8] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194-203, 2001.

[9] M. Haass et al., "Saliency estimation for advanced imaging scenes using pixel statistics," presented at European Conference on Eye Movements 2015, Vienna, Austria, 2015.

[10] L. Matzen et al, "Effects of Professional Visual Search Experience on Domain-General and Domain-Specific Visual Cognition," *Foundations of Augmented Cognition: Lecture Notes in Computer Science*, vol. 9183, pp. 481-491, 2015.

[11] D. S. Wooding, "Fixation maps: Quantifying eye-movement traces," in *Proceedings of ETRA 2002*. ACM, 2002, pp. 31-36.

[12] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, Nov 1998.

[13] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proceedings of Neural Information Processing Systems (NIPS)*, vol. 19, pp. 545-552, 2006.

[14] M. M. Moya, et al, "Superpixel segmentation using multiple SAR image products" in *SPIE Defense+ Security*. International Society for Optics and Photonics, 2014, pp. 90 770R.

[15] M.M. Moya, et al., "Superpixel Classification for Signature Search in Synthetic Aperture Radar Imagery," presented at Conference on Data Analysis (CoDA), Santa Fe, NM, 2014.

[16] X. Ren and J. Malik, "Learning a classification model for segmentation," *Computer Vision, 2003 Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 10–17.

[17] R. Achanta et al., "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 11, pp. 2274–2282, 2012.

[18] J. Tran, "Index for surface coherence (ISC): A method for calculating change susceptibility in SAR change products," in *SPIE Defense + Security, Radar Sensor Technology XX*, Paper 9829-60 [Accepted], 2016.

[19] A. Borji et al, "Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study." *IEEE Transactions on Image Processing,* vol. 22, no. 1, pp. 55-69, 2013.

[20] C. V. Jakowatz et al., *Spotlight-Mode Synthetic Aperture Radar: A Signal Processing Approach*. Boston, MA: Kluwer Academic Publishers, 1996.

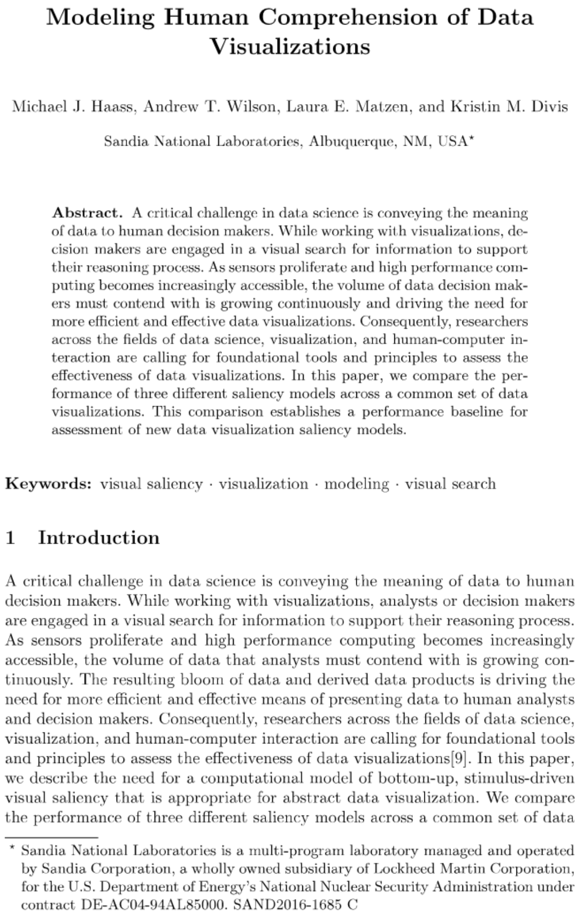
[21] K. M. Simonson. (1998, Aug.). Probabilistic Fusion of ATR Results," Sandia National Laboratories, Albuquerque, NM, SAND Report SAND98-1699.

[22] D. R. Cox and D. V. Hinley, *Theoretical Statistics*, Chapman and Hall, 1974.

Author Biography

Laura Matzen received her BA in linguistics and cognitive science from Rice University (2003) and her PhD in cognitive psychology from the University of Illinois at Urbana-Champaign (2008). Since then she has worked in the Cognitive Science and Systems group at Sandia National Laboratories in Albuquerque, NM. Her work has focused on using cognitive neuroscience methods to understand how humans process and remember information while performing complex reasoning tasks.

APPENDIX B: MODELING HUMAN COMPREHENSION OF DATA VISUALIZATIONS[2]



Modeling Human Comprehension of Data Visualizations

Michael J. Haass, Andrew T. Wilson, Laura E. Matzen, and Kristin M. Divis

Sandia National Laboratories, Albuquerque, NM, USA[*]

Abstract. A critical challenge in data science is conveying the meaning of data to human decision makers. While working with visualizations, decision makers are engaged in a visual search for information to support their reasoning process. As sensors proliferate and high performance computing becomes increasingly accessible, the volume of data decision makers must contend with is growing continuously and driving the need for more efficient and effective data visualizations. Consequently, researchers across the fields of data science, visualization, and human-computer interaction are calling for foundational tools and principles to assess the effectiveness of data visualizations. In this paper, we compare the performance of three different saliency models across a common set of data visualizations. This comparison establishes a performance baseline for assessment of new data visualization saliency models.

Keywords: visual saliency · visualization · modeling · visual search

1 Introduction

A critical challenge in data science is conveying the meaning of data to human decision makers. While working with visualizations, analysts or decision makers are engaged in a visual search for information to support their reasoning process. As sensors proliferate and high performance computing becomes increasingly accessible, the volume of data that analysts must contend with is growing continuously. The resulting bloom of data and derived data products is driving the need for more efficient and effective means of presenting data to human analysts and decision makers. Consequently, researchers across the fields of data science, visualization, and human-computer interaction are calling for foundational tools and principles to assess the effectiveness of data visualizations[9]. In this paper, we describe the need for a computational model of bottom-up, stimulus-driven visual saliency that is appropriate for abstract data visualization. We compare the performance of three different saliency models across a common set of data

[*] Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2016-1685 C

[2] Haass M.J., Wilson A.T., Matzen L.E., Divis K.M. (2016) Modeling Human Comprehension of Data Visualizations. In: Lackey S., Shumaker R. (eds) *Virtual, Augmented and Mixed Reality. VAMR 2016. Lecture Notes in Computer Science*, vol 9740. Springer, Cham.

visualizations to establish a performance baseline for assessment of new data visualization saliency models.

Human visual processing is guided by two parallel processes: bottom-up and top-down visual attention[16]. When viewing an image, a person's eye movements are guided by both the visual properties of the image that capture bottom-up attention (e.g. color, contrast, motion) and top-down processes such as task goals, prior experience, and use of search strategies[8]. Many bottom-up models are based on the neurophysiology of human and primate visual systems[1]. These models construct a number of features from the image data and then highlight differences in the features across multiple scales of image resolution. The chosen features are based on the response of neurons in the visual processing system to certain image characteristics such as luminance, hue, contrast and orientation. Various models have explored the use of different visual features at different scales to predict where humans will look in natural scene imagery.

Maps of bottom-up visual saliency have been valuable tools for studying how people process information in natural scenes, and could also be useful for evaluating the effectiveness of data visualizations. Ideally, the most important information in a data visualization would also have high visual saliency. This evaluation approach has been demonstrated with scene-like data visualizations[12], but it is unclear whether or not it is applicable to abstract data visualizations. In addressing this question, it is important to consider how visual search may differ between natural scene visualizations and abstract data visualizations. For the latter, viewers are engaged in drawing conclusions about causality, efficacy or consequences rather than identifying objects or properties of objects. The visual appearance of their target (information) may not be well defined or known ahead of time. The vast majority, if not all, existing computational models were developed and optimized to predict visual saliency for image-like, or natural, scenes and may not perform as well when applied to abstract data visualizations. In fact one published taxonomy of visual stimuli used in studies of gaze direction lists only three types of stimuli: psychophysics laboratory stimuli, static natural scenes, and dynamic natural scenes[15]. To date, we have been unable to find any published examples of bottom-up saliency models designed explicitly for data visualizations. In the following sections, we compare the performance of three high performing natural scene saliency models across a common set of data visualizations.

2 Method

The MIT Saliency Benchmark[7] is an online source of saliency model performance and datasets. The site scores and reports performance on author-contributed saliency models on datasets where the human fixation positions are not public. This approach prevents model performance inflation due to overfitting of the test dataset. We selected three saliency models, described below, listed on the MIT Saliency Benchmark site that span a range of performance on natural scenes when measured on standard stimuli with a common set of human gaze data. For baseline performance on natural scenes for each model, we used results for the cat2000 data set[4] because it is the most recent (introduced Jan 2015). MATLAB or Python code for each model was downloaded from saliency.mit.edu and saliency maps were constructed with each model on a set of data visualizations. We measured the performance of each model for the data visualizations using the same eight metrics used for the saliency benchmark project. We selected 184 example data visualizations from the Massachusetts (Massive) Visualization Dataset[6] with corresponding eye-movement data[5] from 33 viewers (average 16 viewers per visualization, minimum of 11, maximum of 22). Figure 1 shows an example data visualization and corresponding human fixation map. The MASSVIS samples were selected from infographic blogs, government reports, news media websites and scientific journals.

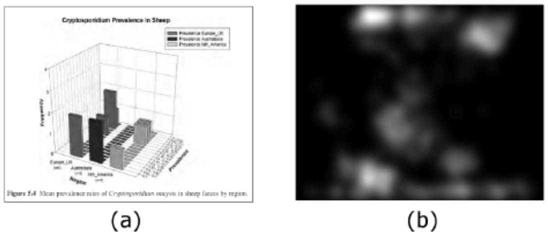


Fig. 1. Example data visualization (a) and human fixation map (b).

2.1 Saliency Models

Itti, Koch and Nieber Numerous saliency prediction models have been developed in recent years, taking a variety of approaches to predict which parts of an image are likely to draw a viewer's attention. Several of these approaches involve the creation of feature maps that are weighted, combined, and filtered to produce a visual saliency map. The most prominent of these models, the Itti, Koch and Niebur model[11], is based on the properties of the human visual system. The model detects changes in low-level features such as color, intensity, and orientation at varying spatial scales. It then weights those features and uses an iterative spatial competition process to create feature maps that are then summed to produce the saliency map. More recently, other researchers have developed new approaches to create saliency maps. When compared using the MIT Saliency Benchmark, two visual saliency models that consistently perform well with images of natural scenes are the Boolean Map based Saliency model[20, 21] and the Ensembles of Deep Networks model[19].

Boolean Map based Saliency The Boolean Map based Saliency model (BMS) [20] creates a set of Boolean maps to characterize images. It relies on the Gestalt principle of figure-ground segregation and the idea that visual attention will be drawn to the figures in an image rather than the background. The model randomly thresholds an image's feature maps, such as the color map, to generate a set of Boolean maps. For each Boolean map, the model uses the feature of surroundedness[21] (a connected region with a closed outer contour) to identify figures within the image and to create an attention map. The attention maps are then normalized and combined to form the full-resolution attention map. This approach differs from many other saliency models because it utilizes scale-invariant information about the topological structure of the images. It does not use multi-scale processing, center-surround filtering, or statistical analysis of features. Thus, it is a relatively simple model that focuses on identifying figures within images.

Ensebles of Deep Networks Like the classic Itti and Koch model, the ensembles of Deep Networks (eDN) model is hierarchical with operations that are based on the known mechanisms of the human visual cortex. However, rather than hand-selecting visual features of interest, a guided search procedure is used to optimize the model for identifying salient features. In other words, the saliency prediction task is a supervised learning problem in which the model is optimized for predicting where humans will look in natural scenes. Multiple high-performing models are identified and the combination of the models is optimized. Center bias and Gaussian smoothing are used to create the final saliency maps from the model outputs. For this comparison, the eDN model coefficients provided by Vig et al., learned using natural scene stimuli rather than data visualizations, were used to illustrate the difference in feature sensitivity across the two stimuli types. Future comparisons of learned model coefficients across the stimuli types could inform the development of saliency models for data visualizations. Figure 2 shows examples of each saliency model applied to the data visualization shown in Fig. 1.

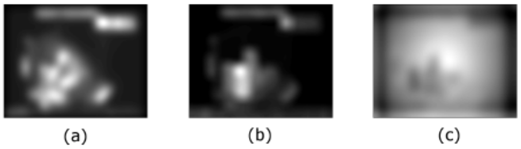


Fig. 2. Example saliency maps, (a) Itti, (b) BMS, (c) eDN, for data visualization shown in Fig. 1.

2.2 Comparison Metrics

Many different metrics have been proposed for comparing saliency and fixation maps. Riche et al. provide a thorough review and taxonomy of published comparison metrics[17]. The authors use a two-dimensional taxonomy to organize the various metrics. Along one dimension, they categorize the metrics as "value-based," "location-based" or "distribution-based." Along the other dimension, they categorize the metrics as "common," "hybrid" or "specific." Metrics categorized as common are generalized and were not originally designed for saliency comparisons. Metrics categorized as hybrid are adapted from other fields to work with saliency and fixation data. Metrics categorized as specific were developed directly for application to saliency comparisons. In order to compare model performance on natural scenes and data visualizations, we elected to use the eight comparison metrics used by the MIT Saliency Benchmark project. Of the eight metrics, one was value-based, three were location-based, and four are distribution-based, as described in more detail below.

Value-based Metric The normalized scanpath saliency metric (NSS)[2] first standardizes saliency values to have zero mean and unit standard deviation, then computes the average saliency value at human fixation locations. When NSS is greater than one, the saliency map exhibits significantly higher values at fixation locations compared to other locations.

Location-based Metrics Three of the comparison metrics are based on the receiver-operator characteristic (ROC). For these metrics, the human gaze positions are considered positive examples and all other points are considered negative examples. The saliency map is treated as binary classifier to separate the positive and negative example sets at various thresholds and the area under the resulting ROC curve (AUC) is computed. As the saliency map and fixation map become more similar, AUC values approach one. Random chance agreement results in an AUC value of 0.5. For all three implementations, the true positive rate is the proportion of saliency values above the threshold at all fixation locations. For the AUC-Judd implementation the false positive rate is the proportion saliency values above the threshold at non-fixated locations and the thresholds are sampled from the saliency map values[17]. For the AUC-Borji implementation, the false positive rate is based on saliency values sampled uniformly from all image pixels and the thresholds are sampled with a fixed stepsize[3]. For the shuffled AUC implementation, the false positive rate is based on saliency values sampled uniformly from fixation locations on a random set of other images[22, 3].

Distribution-based Metrics The similarity score (SIM) is a histogram intersection measure. Each distribution is scaled so that its sum is one. Similarity is the sum of the minimum value between the two scaled distributions at each point. When SIM equals one, the distributions are the same and when SIM equals zero, there is no overlap between the two distributions. The earth mover's distance (EMD)[18] is based on the minimal cost to transform one distribution (the saliency map) into the other distribution (the fixation map). Smaller values of EMD represent better agreement between the saliency map and the fixation map and when EMD equals zero, the two distributions are identical. The linear correlation coefficient (CC) is a measure of the linear relationship between a fixation map and a saliency map[2]. When CC is close to one, the linear relationship between the saliency map and the fixation map is nearly perfect. The KullbackLeibler divergence (KL)[10] is a measure of the information lost when the saliency map is used to approximate the fixation map. KL ranges from zero, when the two maps are identical, to infinity.

3 Experimental Results

Figure 3 shows the performance of the three models on the natural scenes and data visualizations. The results are displayed in the form of a percent difference score that is negative when the models performed better on natural scenes and positive when the models performed better on data visualizations. The corresponding numerical values are shown in Table 1. Table 2 shows the effect size, using Glass's delta across natural scenes and data visualization.

Generally, the models had poorer performance for data visualizations than for natural scenes. All three models performed worse on visualizations than on natural scenes as measured by four of the eight metrics: the value-based metric NSS, two location-based metrics, AUC-Judd and AUC-Borji, and the distribution-based metric EMD. For these metrics, the effect sizes were largest for the BMS and eDN models. The performance of the eDN model was not significantly different for visualizations and natural scenes when measured by the location-based metric sAUC and the distribution-based metric SIM. Similarly, the performance of the Itti model was not significantly different for visualizations and natural scenes when measured by the distribution-based metric CC. However, for the distribution-based metric KL, both the Itti and eDN models performed significantly better for data visualizations than natural scenes. This is consistent with the finding of Riche et al.[17] that the KL metric is quite different from the other metrics. Because the KL metric does not take absolute location into account, but considers only the statistical distribution of the map, two maps having similar distributions can have very different location properties. The performance of the BMS model was not significantly different for visualizations and natural scenes when assessed by the KL metric. For this metric, the effect size was largest for the Itti model followed by the eDN model, while the effect size for the BMS model was close to zero. Of note, for the metrics where the performance of all three models was significantly different between visualizations and natural scenes, the Itti model performed better on visualizations than either the BMS model or the eDN model. This is contrary to the general trend in performance on natural scenes for these metrics where eDN is the best performing saliency model.

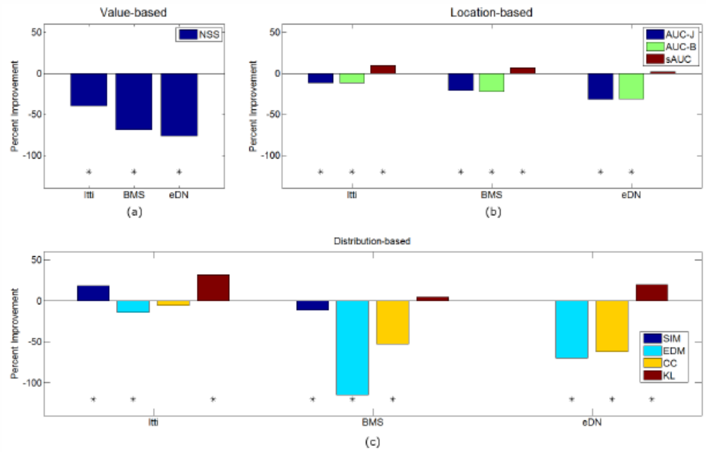


Fig. 3. Model Comparison Across Stimuli Type and Metric. (a) Value-based metric, (b) Location-based Metrics, (c) Distrbution-based Metrics. Results are displayed in the form of a difference score that is negative when the models performed better on natural scenes and positive when the models performed better on data visualizations.

Table 1. Model Comparison Across Stimulus Type. First value in each pair is sample mean; second value is standard error of the mean (SEM). Bold font indicates significant differences between mean values for natural scenes and visualizations ($p < 0.05$).

	Itti		BMS		eDN	
	Nat.	Vis.	Nat.	Vis.	Nat.	Vis.
AUC-J.	**0.77 ± 0.002**	**0.68 ± 0.006**	**0.85 ± 0.001**	**0.67 ± 0.006**	**0.85 ± 0.001**	**0.58 ± 0.009**
SIM	**0.48 ± 0.002**	**0.57 ± 0.006**	**0.61 ± 0.002**	**0.54 ± 0.005**	0.52 ± 0.002	0.52 ± 0.005
EMD	**3.44 ± 0.016**	**3.92 ± 0.11**	**1.95 ± 0.013**	**4.19 ± 0.12**	**2.64 ± 0.013**	**4.48 ± 0.12**
AUC-B.	**0.76 ± 0.002**	**0.67 ± 0.006**	**0.84 ± 0.001**	**0.65 ± 0.006**	**0.84 ± 0.001**	**0.58 ± 0.009**
sAUC	**0.59 ± 0.002**	**0.64 ± 0.007**	**0.59 ± 0.002**	**0.63 ± 0.006**	0.55 ± 0.002	0.56 ± 0.009
CC	0.42 ± 0.004	0.40 ± 0.017	**0.67 ± 0.002**	**0.32 ± 0.014**	**0.54 ± 0.002**	**0.20 ± 0.020**
NSS	**1.06 ± 0.012**	**0.64 ± 0.030**	**1.67 ± 0.012**	**0.52 ± 0.025**	**1.30 ± 0.006**	**0.30 ± 0.032**
KL	**0.92 ± 0.006**	**0.63 ± 0.019**	0.83 ± 0.012	0.79 ± 0.021	**0.97 ± 0.006**	**0.78 ± 0.018**

4 Discussion and Conclusion

The visualizations used in this comparison study are all highly curated, employing text and graphic design principles to help viewers identify the most important results. The Itti model may perform best on these data visualizations because of its close ties to the human visual processing system, while other models have

Table 2. Glass's Delta Effect Size for Model Comparison Across Stimulus Type. Bold font indicates significant differences between mean values for natural scenes and visualizations ($p < 0.05$). For normalization of Glass's delta, the natural scenes were treated as the control group.

	AUC-J.	SIM	EMD	AUC-B.	sAUC	CC	NSS	KL
Itti	**−0.98**	**1.23**	**0.66**	**−0.98**	**0.69**	−0.14	**−0.79**	**−1.16**
BMS	**−3.58**	**−1.00**	**3.79**	**−3.77**	**0.58**	**−3.21**	**−2.09**	−0.07
eDN	**−5.34**	−0.04	**3.07**	**−5.18**	0.14	**−4.22**	**−3.43**	**−0.67**

been designed and optimized for natural scenes, placing less emphasis on faithful representation of neural processes. The natural scene models may also under perform on data visualizations, since many graphical elements used in visualization have smaller spatial extent than objects that typically appear in natural scenes. The finer resolution graphical elements result in higher frequency components to which natural scene models maybe insensitive. Another factor that may limit the applicability of natural scene models is the use of text in data visualizations. Text plays a significant role in human attentional allocation and the resulting direction of eye movements. The process of reading text in a visualization would result in a higher density of fixations around text elements. Future work should leverage a taxonomy of visualization elements such as the one described in Munzner's book[13]. Our future research will focus on data visualization techniques for two-dimensional representation of high-dimensional data.

This comparison study has established a baseline that can be used to assess the performance of new saliency models for data visualizations. The current trend towards better model performance on natural scenes seems to come at the expense of performance on data visualizations. This inverse relationship between model performance on natural scenes and on data visualizations supports our position that new saliency models are needed to aid development of generalized theories of visual search for data visualizations. In future work, we will expand on existing models of visual saliency to address these issues and investigate the role of top-down visual attention in viewers' navigation of abstract data visualizations. Developing general models of top-down sense-making has proven to be quite difficult[14]. Knowledge elicitation techniques have been used to identify top-down goals and strategies and the resulting influence on eye movements. Other approaches have applied machine learning techniques to eye movement data collected as experts perform a given task. The resulting models can predict expert attention allocation for new stimuli, but it is often difficult to use these models to understand why experts allocate attention to certain content and not to other content. Because of this difficulty, we advocate the combination of computational models of bottom-up saliency with empirical studies of eye movements to identify tacit sense-making strategies.

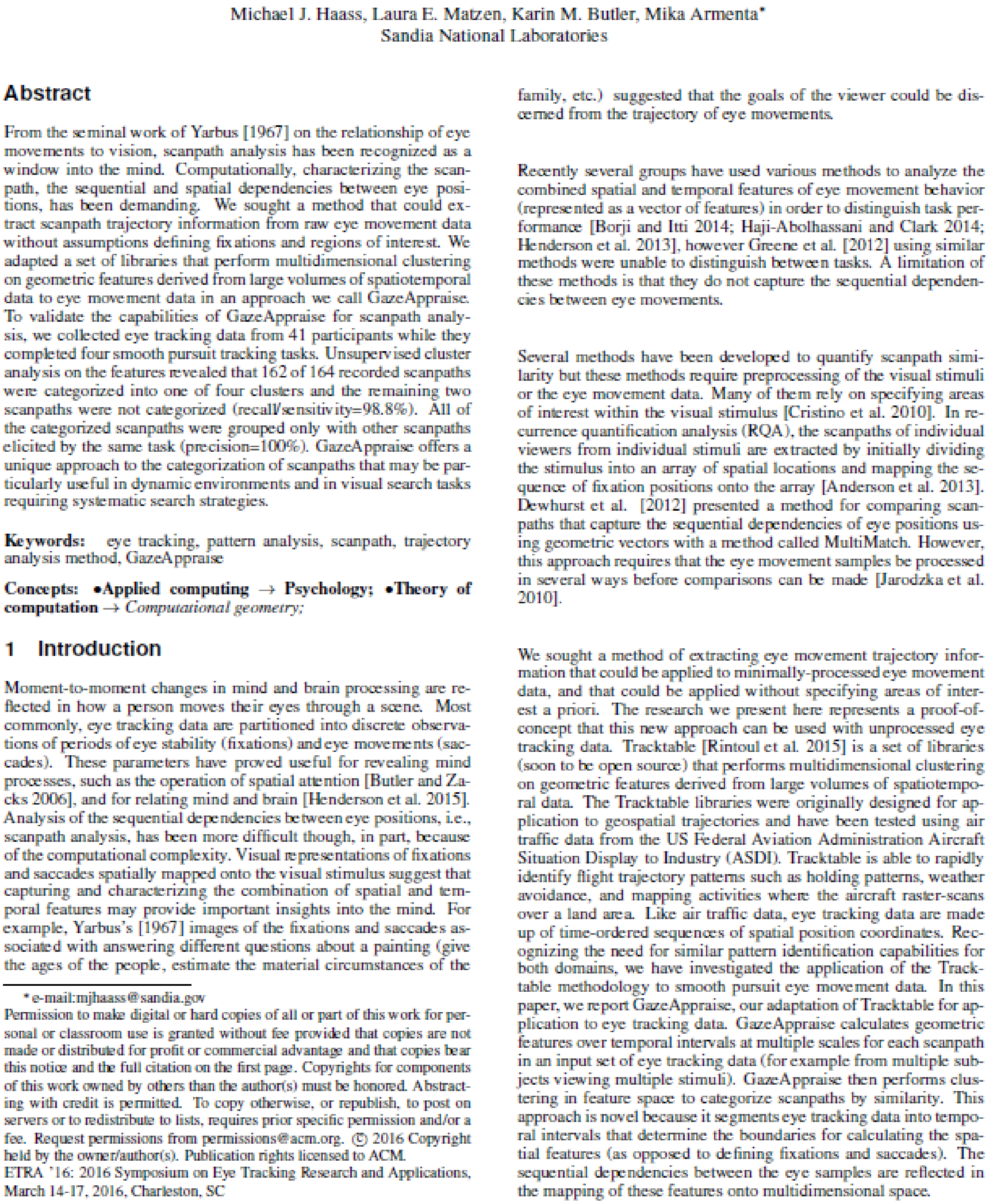
As this work progresses, we will also explore the role of expertise in visual processing of data visualizations. Expertise is a crucial factor in top-down visual attention, and its impact may be even greater with abstract visualizations, where users cannot rely on their prior experience with real-world scenes to guide their search. Visual search tasks using abstract data visualizations can be contrasted with visual search tasks in complex decision making domains. For example, airport luggage screeners search x-ray imagery for prohibited items. In this domain, as in many abstract visualizations, the visual appearance of the target is often not known in advance and furthermore the target may be obscured by overlapping items. However, the users' knowledge about the image features may be quite different. Luggage screening personnel have extensive training and experience in how to search through images, but may have little expertise on the physics of the image formation process. In contrast, experts such as scientists and engineers who work with abstract data are likely to have very deep knowledge of the physical properties driving the content of visualizations. These differences should be considered as top-down factors are identified.

References

1. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. Ieee Transactions on Pattern Analysis and Machine Intelligence (2013)
2. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing (2013)
3. Borji, A., Tavakoli, H.R., Sihite, D.N., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction. IEEE International Conference on Computer Vision (ICCV) (2013)
4. Borji, A., Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. CVPR 2015 workshop on "Future of Datasets" (2015), arXiv preprint arXiv:1505.03581
5. Borkin, M., Bylinskii, Z., Kim, N., C.M., B., Yeh, C., Borkin, D., Pfister, H., Oliva, A.: Beyond memorability: Visualization recognition and recall. IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis) (2015)
6. Borkin, M., Bylinskii, Z., Krzysztof, G., Kim, N., Oliva, A.and Pfister, H.: Massachusetts (massive) visualization dataset, massvis.mit.edu
7. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: Mit saliency benchmark
8. Connor, C.E., Egeth, H.E., Yantis, S.: Visual attention: Bottom-up versus top-down. Current Biology 14(19) (2004)
9. Green, T.M., Ribarsky, W., Fisher, B.: Building and applying a human cognition model for visual analytics. Information Visualization (2009)
10. Itti, L., Baldi, P.: A principled approach to detecting surprising events in video. IEEE Computer Society Conference on Computer Vision and pattern Recognition (CVPR) (2005)
11. Itti, L., Koch, C.and Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Ieee Transactions on Pattern Analysis and Machine Intelligence (1998)
12. Matzen, L.E., Haass, M.J., Tran, J., McNamara, L.A.: Using eye tracking metrics and visual saliency maps to assess image utility. Paper presented at the IS and T International Symposium on Electronic Imaging 2016, Human Vision in Electronic Imaging, San Francisco, CA, USA

13. Munzner, T.: Visualization Analysis and Design. CRC Press (2014)
14. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Research (2005)
15. Peters, R.J., Itti, L.: Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
16. Pinto, Y., van der Leij, A.R., Sligte, I.G., Lamme, V.A.F., Scholte, H.S.: Bottom-up and top-down attention are independent. Journal of Vision 13(3) (2013)
17. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., Dutoit, T.: Saliency and human fixations: State-of-the-art and study of comparison metrics. IEEE International Conference on Computer Vision (ICCV) (2013)
18. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision (2000)
19. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. IEEE Computer Vision and Pattern Recognition (CVPR) (2014)
20. Zhang, J., Sclaroff, S.: Saliency detection: A boolean map approach. Proc. Of the IEEE International Conference on Computer Vision (ICCV) (2013)
21. Zhang, J., Sclaroff, S.: Exploiting surroundedness for saliency detection: A boolean map approach. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2015)
22. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision (2008)

APPENDIX C: A NEW METHOD FOR CATEGORIZING SCANPATHS FROM EYE TRACKING DATA[3]



Michael J. Haass, Laura E. Matzen, Karin M. Butler, Mika Armenta*
Sandia National Laboratories

Abstract

From the seminal work of Yarbus [1967] on the relationship of eye movements to vision, scanpath analysis has been recognized as a window into the mind. Computationally, characterizing the scanpath, the sequential and spatial dependencies between eye positions, has been demanding. We sought a method that could extract scanpath trajectory information from raw eye movement data without assumptions defining fixations and regions of interest. We adapted a set of libraries that perform multidimensional clustering on geometric features derived from large volumes of spatiotemporal data to eye movement data in an approach we call GazeAppraise. To validate the capabilities of GazeAppraise for scanpath analysis, we collected eye tracking data from 41 participants while they completed four smooth pursuit tracking tasks. Unsupervised cluster analysis on the features revealed that 162 of 164 recorded scanpaths were categorized into one of four clusters and the remaining two scanpaths were not categorized (recall/sensitivity=98.8%). All of the categorized scanpaths were grouped only with other scanpaths elicited by the same task (precision=100%). GazeAppraise offers a unique approach to the categorization of scanpaths that may be particularly useful in dynamic environments and in visual search tasks requiring systematic search strategies.

Keywords: eye tracking, pattern analysis, scanpath, trajectory analysis method, GazeAppraise

Concepts: •**Applied computing** → **Psychology;** •**Theory of computation** → *Computational geometry;*

1 Introduction

Moment-to-moment changes in mind and brain processing are reflected in how a person moves their eyes through a scene. Most commonly, eye tracking data are partitioned into discrete observations of periods of eye stability (fixations) and eye movements (saccades). These parameters have proved useful for revealing mind processes, such as the operation of spatial attention [Butler and Zacks 2006], and for relating mind and brain [Henderson et al. 2015]. Analysis of the sequential dependencies between eye positions, i.e., scanpath analysis, has been more difficult though, in part, because of the computational complexity. Visual representations of fixations and saccades spatially mapped onto the visual stimulus suggest that capturing and characterizing the combination of spatial and temporal features may provide important insights into the mind. For example, Yarbus's [1967] images of the fixations and saccades associated with answering different questions about a painting (give the ages of the people, estimate the material circumstances of the family, etc.) suggested that the goals of the viewer could be discerned from the trajectory of eye movements.

Recently several groups have used various methods to analyze the combined spatial and temporal features of eye movement behavior (represented as a vector of features) in order to distinguish task performance [Borji and Itti 2014; Haji-Abolhassani and Clark 2014; Henderson et al. 2013], however Greene et al. [2012] using similar methods were unable to distinguish between tasks. A limitation of these methods is that they do not capture the sequential dependencies between eye movements.

Several methods have been developed to quantify scanpath similarity but these methods require preprocessing of the visual stimuli or the eye movement data. Many of them rely on specifying areas of interest within the visual stimulus [Cristino et al. 2010]. In recurrence quantification analysis (RQA), the scanpaths of individual viewers from individual stimuli are extracted by initially dividing the stimulus into an array of spatial locations and mapping the sequence of fixation positions onto the array [Anderson et al. 2013]. Dewhurst et al. [2012] presented a method for comparing scanpaths that capture the sequential dependencies of eye positions using geometric vectors with a method called MultiMatch. However, this approach requires that the eye movement samples be processed in several ways before comparisons can be made [Jarodzka et al. 2010].

We sought a method of extracting eye movement trajectory information that could be applied to minimally-processed eye movement data, and that could be applied without specifying areas of interest a priori. The research we present here represents a proof-of-concept that this new approach can be used with unprocessed eye tracking data. Tracktable [Rintoul et al. 2015] is a set of libraries (soon to be open source) that performs multidimensional clustering on geometric features derived from large volumes of spatiotemporal data. The Tracktable libraries were originally designed for application to geospatial trajectories and have been tested using air traffic data from the US Federal Aviation Administration Aircraft Situation Display to Industry (ASDI). Tracktable is able to rapidly identify flight trajectory patterns such as holding patterns, weather avoidance, and mapping activities where the aircraft raster-scans over a land area. Like air traffic data, eye tracking data are made up of time-ordered sequences of spatial position coordinates. Recognizing the need for similar pattern identification capabilities for both domains, we have investigated the application of the Tracktable methodology to smooth pursuit eye movement data. In this paper, we report GazeAppraise, our adaptation of Tracktable for application to eye tracking data. GazeAppraise calculates geometric features over temporal intervals at multiple scales for each scanpath in an input set of eye tracking data (for example from multiple subjects viewing multiple stimuli). GazeAppraise then performs clustering in feature space to categorize scanpaths by similarity. This approach is novel because it segments eye tracking data into temporal intervals that determine the boundaries for calculating the spatial features (as opposed to defining fixations and saccades). The sequential dependencies between the eye samples are reflected in the mapping of these features onto multidimensional space.

* e-mail:mjhaass@sandia.gov

ETRA '16: 2016 Symposium on Eye Tracking Research and Applications, March 14-17, 2016, Charleston, SC

[3] Haass, M. J., Matzen, L. E., Butler, K. M., & Armenta, M. (2016, March). A new method for categorizing scanpaths from eye tracking data. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 35-38). ACM.

2 Method

2.1 Participants

Forty-one employees (Males = 22, Females = 16, 3 participants did not self-identify; Age: M = 27.0, SD = 11.5, Range = 17 to 65 years) were recruited from Sandia National Laboratories via email messages distributed to members of the workforce. Participants were paid their typical wage for their time spent participating in this study.

2.2 Apparatus

Eye movements were tracked using Seeing Machines FOVIO running at 60 Hz. The FOVIO was interfaced with EyeWorks Record 3.12 software running on a DELL Precision T3600 and using the Windows 7 operating systems on an Intel Xeon CPU E5-1603 0 @ 2.80 GHz with 8 GB of RAM. Movie files of a moving dot were presented using a script created in EyeWorks Design 3.12. All stimuli were presented on a DELL 19" LCD monitor set at a resolution of 1280 × 1024.

2.3 Materials

The stimuli consisted of four movie files in .avi format in which a white dot (22 × 20 pixels) moved across a black background. The four stimuli were created such that the white dot entered each quadrant of the visual display and so that some of the stimuli had similar curving geometric forms. The shapes traced by the white dot included a star, an S, an O, and a swirl starting from the center of the screen spiraling out. The video dimensions were 1024 × 640 and the movie files were 23, 18, 14, and 14 seconds in length, respectively. At an average viewing distance of 78 cm the dot moved at 6.0 degree of visual angle/sec, a speed that would allow participants to use smooth pursuit eye movements to track the dot. During stimulus presentation each video was preceded by a white fixation cross of 87 × 93 pixels on a black background presented for 2 seconds.

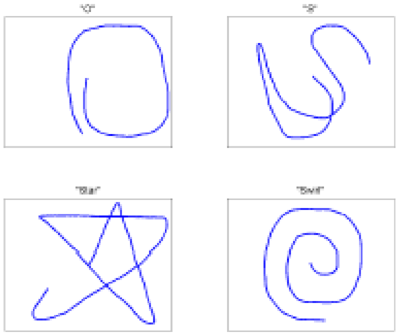


Figure 1: *Four shapes traced by white dot in the smooth pursuit task.*

2.4 Procedure

This study was approved by the Sandia National Laboratories Human Subject Review board. Informed consent was obtained from all participants. Participants were seated in a quiet and darkened room at a distance of 54 to 92 cm from the monitor. Before beginning the eye tracking tasks, the FOVIO was calibrated using a five-point calibration screen. Stimulus presentation was self-paced. Participants were instructed to look at the fixation cross when it appeared and then to follow the white dot as it moved across the screen. The 41 participants generated 164 scanpaths.

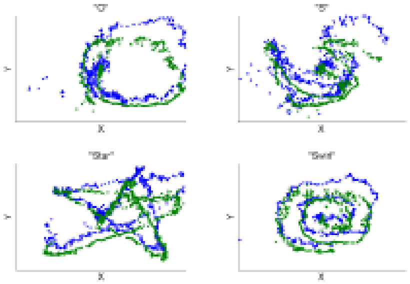


Figure 2: *Sample scanpaths from two randomly chosen subjects for each of four shapes used in smooth pursuit task.*

3 GazeAppraise for Scanpath Analysis

The 164 scanpaths consisting of the x and y position of each sample recorded at 60 Hz were processed using GazeAppraise. In our analysis, we chose 4 temporal scales, resulting in 10 temporal intervals: (1) the entire scanpath, (2 - 3) the first and second halves of the scanpath, (4 - 6) thirds of the scanpath and (7 - 10) quarters of the scanpath. Note that the total number of temporal intervals, T, for the number of temporal scales, n, follows the triangle number series,

$$T_n = \frac{n(n+1)}{2}.$$

We began with 4 temporal scales that had been shown in previous work to minimize computational complexity while providing sufficient resolution to differentiate aircraft trajectories. We found this choice of temporal scales to also be effective in this application to eye movement patterns. Following the Tracktable method of Rintoul et al. [2015] let $SP(t)(t \in [0\ 1])$ represent the entire scanpath, then the set of scanpath temporal intervals is:

(1) $SP(t)(t \in [0\ 1])$

(2 - 3) $SP(t)(t \in [0\ \frac{1}{2}])$ and $SP(t)(t \in [\frac{1}{2}\ 1])$

(4 - 6) $SP(t)(t \in [0\ \frac{1}{3}])$, $SP(t)(t \in [\frac{1}{3}\ \frac{2}{3}])$ and $SP(t)(t \in [\frac{2}{3}\ 1])$

(7 - 10) $SP(t)(t \in [0\ \frac{1}{4}])$, $SP(t)(t \in [\frac{1}{4}\ \frac{1}{2}])$, $SP(t)(t \in [\frac{1}{2}\ \frac{3}{4}])$ and $SP(t)(t \in [\frac{3}{4}\ 1])$

One or more features can be calculated over each temporal interval. For the smooth pursuit task, we calculated a two dimensional feature at each temporal interval: the median x and y position of the gaze. This metric was chosen because it is a robust statistic; it is less sensitive to noise in the eye tracking samples introduced by the specific eye tracking system or study environment conditions (such as subjects free to move in the eye tracker's head box volume). Let $md(SP[t0\ t1])$ be the median x and y location of the scanpath samples contained in the temporal interval $[t0\ t1]$, then the set of 10, two dimensional, features describing the scanpath is:

(1) $md(SP)[0\ 1]$
(2 - 3) $md(SP)[0\ \frac{1}{2}]$ and $md(SP)[\frac{1}{2}\ 1]$
(4 - 6) $md(SP)[0\ \frac{1}{3}], md(SP)[\frac{1}{3}\ \frac{2}{3}]$ and $md(SP)[\frac{2}{3}\ 1]$
(7 - 10) $md(SP)[0\ \frac{1}{4}], md(SP)[\frac{1}{4}\ \frac{1}{2}], md(SP)[\frac{1}{2}\ \frac{3}{4}]$ and $md(SP)[\frac{3}{4}\ 1]$

The median calculation is implemented using the BOOST C++ library (www.boost.org) using a P2 quantile estimation algorithm.

To illustrate the feature calculation process, Figure 3 shows the vertical position (y axis) of the gaze of an ideal viewer versus elapsed time for the star smooth pursuit pattern. For this example, an ideal viewer would produce gaze coordinates that exactly match those of the stimulus dot as it moves over the screen. The horizontal lines at the top of the figure show three of the temporal scales used to calculate the median gaze y location at each of six temporal intervals. The triangles indicate the median gaze location feature value calculated at each of the corresponding temporal intervals. The 10, two-dimensional features calculated for the smooth pursuit tasks were represented in 20-dimensional space. Unsupervised cluster analysis was performed using a scale-insensitive approach based on the well-known density based spatial clustering algorithm, DBSCAN [Ester et al. 1996]. For density based clustering, the total number of clusters does not need to be specified a priori. Instead, two intuitive parameters, the minimum number of members required to form a cluster (minPts) and the neighborhood radius (Eps), influence cluster identification. We set minPts equal to 10 (˜1/4th of the total number of subjects) and Eps equal to 200 pixels (˜1/6th of the full horizontal screen width). To illustrate the multidimensionality of these metrics, Figure 4 displays a subset of data. It depicts the x dimension data from 3 features calculated for the four different stimuli and each participant. Color indicates cluster membership identified by GazeAppraise. This figure also illustrates the need for a density based clustering algorithm rather than other approaches such as k-nearest-neighbors. Clusters are clearly present in the feature set, but the complexity of the cluster boundaries increases with increasing dimensionality of the feature space.

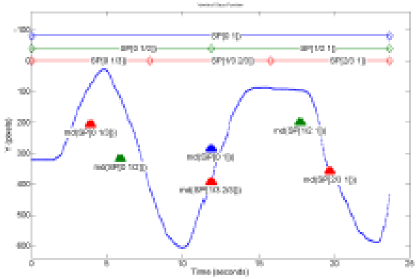


Figure 3: *Vertical position of the gaze of an ideal viewer versus elapsed time for the star stimulus. Three upper, horizontal lines show temporal intervals used to calculate features. Triangles show feature values at each temporal interval.*

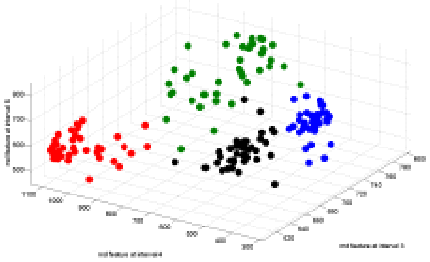


Figure 4: *Three dimensional view of feature data used in unsupervised clustering. Color indicates cluster membership identified by GazeAppraise.*

4 Results

Table 1 presents the results of applying the GazeAppraise algorithm to the 164 scanpaths. Unsupervised cluster analysis revealed that 162 of the scanpaths were categorized into one of four clusters and the remaining two scanpaths were considered outliers and not categorized, resulting in a recall/sensitivity score of 98.8%. All of the categorized scanpaths were grouped only with other scanpaths elicited by the same task for a precision = 100%.

Table 1: *Number of scanpaths assigned to each cluster in unsupervised clustering of 164 scanpaths.*

	Stimuli			
Cluster	O	S	Star	Swirl
1	40			
2	0	41		
3	0	0	40	
4	0	0	0	41
Outlier	1	0	1	0

5 Discussion

When GazeAppraise was applied to unprocessed eye tracking data to extract spatiotemporal features, the resulting multidimensional data were clustered into categories that reflected the differences between the original stimuli. This study represents a proof-of-concept; GazeAppraise successfully categorized raw eye tracking samples into distinct scanpaths that reflected the stimulus constraints, but in the absence of stimulus information to constrain the categorization.

One advantage of GazeAppraise is that, unlike previous scanpath analysis techniques (e.g., Multimatch; Dewhurst et al., [2012]), GazeAppraise does not require preprocessing of the eye movement data into fixations and saccades. Calculating these parameters from eye movement data requires assumptions that define which samples are part of fixations and which samples are from saccades. Parsing the eye movement record into discrete units (fixations and saccades) becomes more complex in dynamic environments where fixating a visual stimulus may require smooth pursuit eye movements, or when saccades may not be required to "fixate" a new object because the visual scene has changed.

Another advantage of GazeAppraise is that the approach does not require defining areas of interest or arrays of spatial locations a priori. Rather it classifies similar scanpath shapes together in the absence of stimulus information or knowledge. This ability is important because as visual stimuli become more cluttered and dynamic the requirement of characterizing the spatial location of important information, or a relevant spatial array, becomes more onerous. Indeed, GazeAppraise can categorize the spatial dependencies between eye movement samples in the absence of a visual stimulus, thus providing a means of characterizing eye movements that are related to visual imagery and mindwandering.

In their guide to eyetracking, Holmqvist and colleagues identified several scanpath comparisons that could be useful [Holmqvist et al. 2011]. Of the seven listed, GazeAppraise has the potential to address four of them: (1) overall shape comparison, (2) similar shape that differs in scale, (3) similarity in position but reversal of order, and (4) differences in the speed of execution of a scanpath.

In this paper, we demonstrated that GazeAppraise can categorize scanpaths from raw eye tracking data, even when those data include samples collected with variations in calibration precision, tracking consistency, and viewer performance. Future work will need to explore how much and in what ways shapes can differ but still be categorized together. Similarly, scaling algorithms applied to the calculation of x and y features could allow similarly-shaped scan paths that differ in scale to be clustered together, while representation of the reversed order of a set of positions at different temporal scales would also be relatively straightforward to implement.

Although not tested here, it can be mathematically shown that, GazeAppraise will cluster together similar scanpaths that vary in temporal duration when the differences in time are distributed evenly across the eye movement samples relative to the duration of the scanpath. It remains to be demonstrated how robust GazeAppraise is to uneven distribution of these temporal differences across a viewing event. For example, it is expected that there would be more temporal variation across individuals in eye movement samples collected during cognitively guided viewing than during saliency guided viewing.

Although this application of GazeAppraise used the median x and y position as features, the metrics used for each feature are flexible. In fact, each feature can have a different units scale, i.e. one feature measured in degrees of visual angle, another measured in milliseconds and another measured in pixels. Thus, features can be any quantity calculable from the eye tracking samples in each temporal interval. Other features that may be useful for scanpath categorization include, but are not limited to, mean and variance of point-to-point distances, mean nearest neighbor distance (randomness of points), total length of scanpath, area and centroid of the convex hull encompassing scanpath points, etc. For example, metrics based on point-to-point distances would implicitly encode the proportion of fixation to saccade activity over the temporal interval. Total scanpath length could measure the amount of the visual display that was viewed which may be important for assessing systematic search processes like visual inspection. Convex hull metrics could measure the amount of the peripheral visual display that is viewed.

The application of GazeAppraise to eye movement analysis is nascent; the eye tracking samples were collected under highly constrained viewing conditions (smooth pursuit eye movements constrained by the stimulus characteristics) not typical of everyday eye movement patterns. It remains to be demonstrated that more typical eye movement trajectories with fixations and saccades, that are influenced to a greater extent by top-down processes, can be categorized. The contribution of this research is to demonstrate the application of a new set of spatiotemporal trajectory libraries to raw eye tracking data, an application we refer to as GazeAppraise. Categorization of eye tracking data collected while viewing four different, but constraining, stimuli was highly successful. Future work will validate the usefulness of this approach by applying the algorithm to eye tracking data from systematic search tasks.

Acknowledgements

Our thanks go to Danny Rintoul and Andy Wilson of Sandia National Laboratories, Albuquerque, NM, USA, for their assistance adapting Tracktable to eye tracking data.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

ANDERSON, N. C., BISCHOF, W. F., LAIDLAW, K. E., RISKO, E. F., AND KINGSTONE, A. 2013. Recurrence quantification analysis of eye movements. *Behav Res Methods 45*, 3, 842–856.

BORJI, A., AND ITTI, L. 2014. Defending yarbus: eye movements reveal observers' task. *Journal of Vision 14*, 3.

BUTLER, K. M., AND ZACKS, R. T. 2006. Age deficits in the control of prepotent responses: evidence for an inhibitory decline. *Psychol Aging 21*, 3, 638–643.

CRISTINO, F., MATHOT, S., THEEUWES, J., AND GILCHRIST, I. D. 2010. Scanmatch: a novel method for comparing fixation se-quences. *Behav Res Methods 42*, 3, 692–700.

DEWHURST, R., NYSTROM, M., JARODZKA, H., FOULSHAM, T., JOHANSSON, R., AND HOLMQVIST, K. 2012. It depends on how you look at it: scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behav Res Methods 44*, 4, 1079–1100.

ESTER, M., KRIEGEL, H., SANDER, J., AND XU, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96 Proceedings. Second International Conference on Knowledge Discovery and Data Mining*, 226–231.

GREENE, M. R., LIU, T., AND WOLFE, J. M. 2012. Reconsidering yarbus: a failure to predict observers' task from eye movement patterns. *Vision Research 62*, 1–8.

HAJI-ABOLHASSANI, A., AND CLARK, J. J. 2014. An inverse yarbus process: predicting observers' task from eye movement patterns. *Vision Research 103*, 127–142.

HENDERSON, J. M., SHINKAREVA, S. V., WANG, J., LUKE, S. G., AND OLEJARCZYK, J. 2013. Predicting cognitive state from eye movements. *PLoS One 8*, 5.

HENDERSON, J. M., CHOI, W., LUKE, S. G., AND DESAI, R. H. 2015. Neural correlates of fixation duration in natural reading: Evidence from fixation-related fmri. *Neuroimage 119*, 390–397.

HOLMQVIST, K., NYSTRÖM, M., ANDERSSON, R., DEWHURST, R., JARODZKA, H., AND VAN DE WEIJER, J. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press.

JARODZKA, H., HOLMVQUIST, K., AND NYSTRÖM, M. 2010. A vector-based, multidimensional scanpath similarity measure. *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, 211–218.

RINTOUL, M., WILSON, A., VALICKA, C., SHEAD, T., RODRIGUEZ CZUCHLEWSKI, K., KEGELMEYER, W., AND NEWTON, B., 2015. Panther: Trajectory analysis. Technical report.

YARBUS, A. L. 1967. Eye movements and vision. *Plenum Press*.

APPENDIX D: PATTERNS OF ATTENTION: HOW DATA VISUALIZATIONS ARE READ[4]

Laura E. Matzen, Michael J. Haass, Kristin M. Divis, Mallory C. Stites

Sandia National Laboratories, Albuquerque, USA
lematze@sandia.gov, mjhaass@sandia.gov, kmdivis@sandia.gov, mcstite@sandia.gov

Abstract. Data visualizations are used to communicate information to people in a wide variety of contexts, but few tools are available to help visualization designers evaluate the effectiveness of their designs. Visual saliency maps that predict which regions of an image are likely to draw the viewer's attention could be a useful evaluation tool, but existing models of visual saliency often make poor predictions for abstract data visualizations. These models do not take into account the importance of features like text in visualizations, which may lead to inaccurate saliency maps. In this paper we use data from two eye tracking experiments to investigate attention to text in data visualizations. The data sets were collected under two different task conditions: a memory task and a free viewing task. Across both tasks, the text elements in the visualizations consistently drew attention, especially during early stages of viewing. These findings highlight the need to incorporate additional features into saliency models that will be applied to visualizations.

Keywords: Data visualizations, text, eye tracking

Introduction

Data visualizations are widely used to convey information, yet it is difficult to evaluate whether or not they are effective. Previous work on graph comprehension has suggested that the effectiveness of a graph depends on the relationships between the visual properties of the graph, the experience and expectations of the user, and the type of information to be extracted from the graph (reviewed in [27]). As such, the recommendations for the "best" way to present as dataset may differ for every new visualization created.

Eye tracking can provide insight into how people comprehend data visualizations. It is a useful measure of where visual attention is being directed, as attention is typically closely linked with gaze location (see [25] for review). Eye tracking measures are divided into fixations (periods of relative stability) and saccades (ballistic movements, during which effectively no new visual information is processed). In general, people tend to spend more time looking at, and make more fixations on, areas of a display that are difficult to process or important to their current task goals [25]. Graph comprehension researchers have devised various metrics to evaluate ease of processing information from graphs. For example, the time to the first fixation in a region is taken as an indicator of how easy the region was to find. The time from landing in a region to making a decision about a graph is taken as an indicator of how easy the information was to process after it was found (see [5] and [12] for discussions of other useful metrics). In this way, eye movement patterns can provide a window into the ongoing cognitive processes taking place as people comprehend data visualizations.

Although eye tracking metrics have the potential to be useful in evaluating the effectiveness of a data visualization in conveying information to a viewer, they must be evaluated within the context of many different factors that affect viewers' eye movement patterns. One factor is the viewer's task, which has a large impact on his or her eye movements. For example, Goldberg and Helfman [12] found more fixations to a graph when viewers subtracted or added data than when they were tasked with simply extracting values. Similarly, Strobel et al. [28] found more fixations to line graphs than bar graphs when users were performing trend analyses. The type of visualization technique used also impacts how users take in the same information, with, for example, more fixations for unfamiliar or difficult visualizations [10,11]. Characteristics of the viewer also influence eye movement behaviors. More experienced users can extract information in less time and may pay attention to different aspects of a visualization than less experienced viewers [21].

To address the diversity of factors that can influence what aspects of a data visualization draw the viewer's attention, it is useful to distinguish between top-down and bottom-up visual attention. Top-down, or goal-oriented, visual attention is driven by the viewer's goals and expectations. Meanwhile, bottom-up visual attention is driven by the physical characteristics of the image, such as color and contrast [9,23]. There are existing models of bottom-up visual attention that use the visual properties of an image to predict which parts of the image will draw a viewer's attention (cf. [16]). These models take an input image and generate a map of visual saliency, where the salient regions are those that are more likely to attract bottom-up visual attention. To assess the ability of the models to predict where people will look, the saliency maps are compared to eye

[4] Matzen, L. E., Haass, M. J., Divis, K. M., & Stites, M. C. (2017, July). Patterns of Attention: How Data Visualizations Are Read. In *International Conference on Augmented Cognition* (pp. 176-191). Springer, Cham.

tracking data collected under free viewing conditions (i.e. the participants view the images for a fixed amount of time with no specific task to complete; [2]).

In prior work, we developed evaluation approaches for data visualizations that incorporate eye tracking data, saliency maps, and sensor phenomenology [21]. We demonstrated that comparing saliency maps to eye tracking data collected from experienced and inexperienced viewers can highlight the differences between features that are highly salient and features that are highly task-relevant. Using saliency maps and eye tracking data in combination was informative for teasing apart which aspects of the data drew viewers' attention from both the bottom-up and top-down perspectives. This information can then be applied to improving the visual representation of the data and to assessing feature detection algorithms.

In subsequent work, we have attempted to extend this general approach from the realm of sensor data into the domain of abstract data visualizations. Predicting what parts of a visualization will draw the user's attention would be a useful first pass at evaluation [26]. However, our work has found that existing saliency maps do not work well for predicting where viewers will look in abstract data visualizations. In Haass et al. [13], we evaluated the ability of multiple models of visual saliency to explain viewing behaviors in natural scenes as well as data visualizations. The models performed well for natural scenes, but they were poor predictors of viewing patterns for abstract data visualizations. Based on comparisons of the saliency maps and fixations, a large part of the discrepancy seems to be due to people attending to text in the data visualizations. The text elements received a high proportion of the viewers' fixations, but were generally not identified as salient in the saliency maps. The visual properties of text are quite different from those of features in natural scenes, so models designed to predict eye movement in scene viewing do not account for the text's influence on the viewer's patterns of attention.

The findings of Haass and colleagues [13] highlight the point that abstract data visualizations are very different from natural scenes – each element was chosen by a designer and is there for a reason. In this way, data visualizations share some commonalities with print ads, which are also comprised of a combination of images and text to convey a message. Eye-tracking techniques have been applied to the print ad literature (see review in [14]), and their findings have largely echoed the graph comprehension literature in showing that the viewer's goals have a huge influence over eye movement guidance. One robust finding is that when viewers are asked to learn about a product or decide on a product to purchase, they tend to look at the text of an ad earlier and for more time—roughly 70% of viewing time—than when they are evaluating an ad for its likeability or effectiveness (in which case viewers show a preference for fixating the images). Readers are also more likely to fixate, and spend more time viewing, ads with large text relative to small text, although the same is not true for photo size. Importantly, the characteristics of eye movements also change when people look at different elements of ads: readers make longer fixation durations and saccades on graphical elements compared to text.

It is worth noting that the graphical elements in ads and data visualizations serve different purposes (display a product versus convey numeric information, respectively), and so different mechanisms might influence viewing patterns for these two visualization types. However, gaining an understanding of the features that drive eye movements in a range of visualizations is an important first step in understanding how viewers allocate their attention between text and graphics during successful comprehension. Uncovering these basic features will help inform models of visual saliency. Our previous work has already shown that simple saliency maps are not sufficient to explain viewing patterns in visualizations [13]. Updating these models to incorporate insights regarding how users allocate their attention between text and graphics might help visualization designers to assess their designs more accurately than models that treat text similarly to graphics.

In the present study, we take a closer look at viewers' attention to text in data visualizations. First, we analyzed eye tracking data collected by Borkin and colleagues [3] in the context of a memory study. While their study included a wide range of visualizations, we selected and analyzed a subset of the data that included frequently-used graph types, such as bar charts and line graphs. We then assessed how much attention participants devoted to different regions of the visualizations, paying particular attention to how attention was allocated to regions that contained text compared to those that did not. The data collected by Borkin et al. [3], henceforth referred to as the MASSVIS data, was collected during a memory study. The parameters of this task are somewhat different from those used in the eye tracking datasets that are commonly used to evaluate visual saliency models. To address this, we collected eye tracking data from a new group of participants who completed a free viewing task for the same subset of the MASSVIS images and an additional set of newly created data visualizations.

Viewing Data Visualizations in a Memory Task

To study how viewers divide their attention between text and graphics in data visualizations, we began with an analysis of a subset of the MASSVIS dataset (http://massvis.mit.edu/). These data were collected during a memory study in which participants viewed images for 10 seconds and were later tested on their memory for the visualizations via recognition and recall tests [3].

For the present analysis, we selected a subset of 35 images from the MASSVIS study. These images represented a variety of commonly used types of data visualizations, all of which contained some combination of text and graphical representations of data. The subset included four area plots, four bar charts, one bubble plot, four column charts (including two double Y-axis plots in which a line graph was overlaid on the column charts), three correlation plots, three line graphs, two map-based visualizations, three network diagrams, three pie charts, and five scatter plots. In addition to these 32 images, we included the three visualizations that had the best match between the eye tracking data and the saliency maps in our prior evaluation of saliency models [13]. These included two infographics and one line graph.

Regions of interest (ROIs) were defined for the stimulus set, dividing the visualizations into the following regions: Title, Data, Data Area, X-Axis, X-Axis Label, Y-Axis, Y-Axis Label, Legend, Data Labels, and Text. For each visualization, the ROIs were marked using GIMP software (www.gimp.org). The ROIs were tightly drawn to the edges of each region.

Scan paths, representing the sequence of fixations across the ROIs for each participant and each visualization where constructed using MATLAB [20]. Fixations were counted as falling within an ROI if their center, defined as the geometric median of all points in the fixation, fell within a 1 degree viewing angle of the ROI, approximating the participants' useful field of view. If the same fixation could be assigned to multiple ROIs, multiple variants of the scan path were generated. However, for the purpose of this analysis, only the first variant was used. A total of 562 scan paths were analyzed, with an average of 16 scan paths from different participants for each visualization. There were an average of 36 fixations per scan path (range 6-51).

Analyses

For each visualization, the number of participants who fixated within each ROI in the visualization at least once was calculated. The average proportion of participants who fixated on an ROI (when present) across all of the visualizations is shown in Table 1. Unsurprisingly, participants nearly always fixated on the data in the visualizations. They were also highly likely to fixate on the title, legend, and data labels, when those ROIs were present.

To determine where the participants allocated their attention in the visualizations, we calculated the proportion of each participant's fixations that fell within each ROI for each visualization. The average proportion of fixations in each ROI is also shown in Table 1. The Data ROI received the highest average proportion of fixations, but this proportion was relatively low. On average, only 27% of the participants' fixations were in the Data ROI, while the Title and Data Labels ROIs received similar proportions of fixations (25% and 26%, respectively).

Table 1. Attention to each ROI in the analysis of the MASSVIS data, including average proportions and (standard deviations).

ROI Name	Number of visualizations containing ROI	Average proportion of participants viewing ROI	Average proportion of fixations to ROI
Title	26	0.94 (0.10)	0.25 (0.10)
Data	35	0.98 (0.05)	0.27 (0.17)
Data Area	21	0.55 (0.26)	0.04 (0.03)
X-Axis	24	0.64 (0.20)	0.05 (0.03)
X-Axis Label	11	0.67 (0.14)	0.06 (0.05)
Y-Axis	24	0.70 (0.22)	0.12 (0.17)
Y-Axis Label	15	0.73 (0.25)	0.10 (0.08)
Legend	23	0.89 (0.15)	0.20 (0.11)
Data Label	15	0.88 (0.22)	0.26 (0.16)
Text	24	0.56 (0.28)	0.07 (0.10)

To test our hypothesis that participants disproportionately pay attention to text in data visualizations, the ROIs were categorized based on whether or not they contained text for each stimulus. For example, the X-Axis ROIs contained text in some visualizations but not in others. For each visualization, we then calculated the proportion of fixations that fell in ROIs containing text, the proportion of fixations to the data and data area, and the proportion of fixations that fell in other ROIs that did not contain text (including graphics, symbols, numbers, etc.). On average across all of the visualizations, 59.9% (SD = 16.1%) of the participants' fixations fell into ROIs containing text relative to 30.0% (SD = 15.6%) of fixations in the data ROIs and 10.1% (SD = 6.6%) of fixations in the other non-text ROIs.

As another measure of how participants weighted the relative importance of each ROI, we assessed how often each ROI was one of the first three ROIs visited by a participant. This was calculated as the proportion of scan paths in which the ROI was one of the first three fixated (for visualizations where that ROI was present). Note that this does not necessarily mean that one of the first three *fixations* in the trial fell in that ROI. For example, if a participant began a trial by fixating four times on the title, then fixating three times on the data, and then fixating once on the legend, then the title, data, and legend would be counted as the first three ROIs visited on that trial. In other words, we assessed the order in which the ROIs were viewed irrespective of the number of fixations in the sequence.

The Title ROI was the most likely to be one of the first three ROIs visited. When the Title ROI was present in a visualization, it was one of the first three visited in 87.8% of the scan paths. The Data ROI was a close second at 83.5%. The proportions were much lower for the other ROIs (51.1% for Data Labels; 39.8% for Legend; 34.7% for the combination of Y-Axis and Y-Axis Labels; 17.0% for the combination of X-Axis and X-Axis Labels; 14.8% for Text). Some of the X- and Y-Axis ROIs contained words (e.g. the names of countries or months) while others were numerical (e.g. years or values). The axis ROIs were subdivided into those that contained text (other than the axis labels) and those that did not. When the X-Axis ROI contained text, it was one of the first three ROIs visited in 48.5% of the scan paths.[5] When the X-Axis ROI did not contain text, it was one of the first three ROIs visited in 12.4% of the scan paths. The difference was even more dramatic for the Y-Axis ROI, which was in the first three ROIs visited in 80.9% of the scan paths when the ROI included text, but only 13.0% of the scan paths when it did not.

To explore the data further, we looked at correlations between the number of words in an ROI and the proportion of fixations in the ROI. If a participant is spending time reading the text in a particular ROI, we would expect to see a high correlation between the number of words and the proportion of fixations. The correlations were significant for the Title ($R^2 = 0.73$, $p < 0.001$), Text ($R^2 = 0.82$, $p < 0.001$), X-Axis Label ($R^2 = 0.69$, $p < 0.02$), and Y-Axis Label ($R^2 = 0.83$, $p < 0.001$) ROIs. For the Legend and Data Label ROIs, which received relatively high proportions of fixations on average, there was not a significant correlation between the number of words and the proportion of fixations (Legend: $R^2 = 0.39$, $p = 0.07$; Data Labels: $R^2 = 0.41$, $p = 0.15$).

The axes themselves provide an interesting opportunity for investigating the effect of text on where viewers spend their time when studying a visualization. As mentioned above, some of the X- and Y-Axis ROIs contained words and others contained only numbers. When the axes contained words, there was a significant correlation between the number of words and the proportion of fixations to the axis (X-Axis: $R^2 = 0.48$, $p < 0.02$; Y-Axis: $R^2 = 0.90$, $p < 0.001$). In contrast, when the X-Axis contained only numerical values, there was no correlation between the number of numerical values and the proportion of fixations ($R^2 = 0.09$, $p = 0.68$). When the Y-Axis contained only numerical values, there was a significant *negative* correlation ($R^2 = -0.46$, $p < 0.03$).

Discussion

The results of our analyses indicate that participants disproportionately viewed regions of the visualizations that contained text in the MASSVIS study. Although the participants did spend time looking at the visualized data, the majority of their fixations were devoted to regions containing text. For some of those regions, including the Title, Text and Axis Label ROIs, significant correlations between the number of fixations and the number of words in the ROIs indicate that participants were spending time reading the text. For other regions, namely the Legend and Data Label ROIs, there was not a significant correlation between the number of fixations and the number of words. These ROIs received relatively high proportions of fixations overall, so the absence of a correlation between the number of words and the proportion of fixations in these regions likely indicates that the participants read the text in those regions but also referred back to them more than once as they studied the visualizations.

Interestingly, the axes of graphs seemed to attract participants' attention when they contained text but not when they contained numbers. Axes containing text were much more likely to be one of the first three ROIs viewed than axes containing only numbers, and for the Y-Axis ROI there was a significant negative correlation between the number of fixations and the number of numerical values along the axis. There are several possible explanations for this pattern, but it seems plausible that numerical axes can be comprehended at a glance, making repeated fixations and revisits unnecessary.

An important point to note is that the MASSVIS eye tracking dataset was collected in the context of a memory study, which may have had a substantial influence on how participants allocated their attention. For example, they may have devoted a lot of attention to the titles of the graphs, thinking that the titles would be easier to remember than the details of the visualized data. To explore the impact of the task on patterns of attention to the visualizations, we conducted a study in which participants viewed data visualizations in a free viewing task.

Viewing Data Visualizations in a Free Viewing Task

When eye tracking datasets are used to assess saliency maps, the participants in the eye tracking studies are typically given a free viewing task. For example, in the widely used MIT Saliency Benchmark eye tracking datasets (http://saliency.mit.edu),

[5] However, there were only two visualizations in this category, with a total of 33 scan paths. The other groupings contained much higher numbers of visualizations and scan paths.

participants completed a free viewing task in which they viewed each image for 5 seconds [2, 6, 17]. In this study, we used the same task and presentation duration to examine eye movement patterns on a larger set of data visualizations and a larger group of participants. Participants viewed the same subset of MASSVIS stimuli that were used in the analysis described above and an additional 27 data visualizations in the context of a larger free viewing experiment.

Method

Participants.

Thirty participants were recruited from students, faculty, and staff in the University of Illinois community (10 males; mean age = 30.53 years, SD = 13.06) and compensated $20 for their time. All participants were tested for color vision deficiencies (24 plate Ishihara Test [15]) and near vision acuity prior to completing the study. Data from an additional five participants was discarded because: they failed the colorblindness and/or acuity tests prior to beginning the experiment (2 participants); the eye tracker failed to successfully capture their eye movements for a significant portion of the experiment (1 participant); they fell asleep for any portion of the experiment (1 participant); or there was a problem with the experimental apparatus (1 participant).

Materials.

Four blocks of images were used in this study, consisting of a total of 108 images. Each image was centered and gray padded to fill the dimensions of the screen.

Two of the blocks consisted of line drawings (30 images) and fractals (16 images) drawn from the MIT Saliency Benchmark CAT2000 dataset [2]. Those blocks are not analyzed in the present study. One block contained thirty-five data visualizations pulled from the MASSVIS dataset [3, 4]. These were the same visualizations as those analyzed in section 2. The final block contained twenty-seven data visualizations that were created specifically for this experiment (3 bar charts, 3 boxplots, 3 bubble graphs, 3 column charts, 3 line plots, 3 parallel coordinates plots, 3 pie charts, 3 scatterplots, and 3 violin plots[6]). These stimuli were selected to represent a variety of common types of data visualizations. To mirror the visualizations in the MASSVIS set, not all of the visualizations contained all of the possible ROIs and the placement of specific ROIs (such as the Legend) varied across visualizations. The newly generated visualizations also differed from the MASSVIS set because they did not contain infographics or additional text, such as text indicating the source of the data.

The order in which the four blocks of images were presented was counterbalanced across participants. Within each block, the stimuli were shown in a random order.

Procedure.

The experiment was completed in a dark room at a nominal viewing distance of 0.8 meters. Stimuli were presented on a large monitor (0.932 x 0.523 meters; 1920 x 1080 pixels) while eye movements were recorded with two Smart Eye Pro cameras. Participants first underwent the standard Smart Eye camera setup procedure and 9-point calibration.

Participants were instructed to view each image as it was presented. Each trial began with a 2-second fixation cross in the center of the screen. The fixation cross was followed by the presentation of an individual image, which was displayed on the screen for 5 seconds.

Analysis.

In the resulting dataset, fixations were defined as samples for which the velocity over the preceding 200 milliseconds (ms) was less than 15 degrees per second. The first fixation in each trial and any fixations with a duration less than 100 ms were dropped from the analysis. For all of the analyses described below, the visualizations pulled from the MASSVIS set and the visualizations created specifically for this experiment are pooled together. A total of 1834 scan paths were included in the analysis. There were an average of 11 fixations per scan path (range 1-19).

As in our earlier analysis, the number of participants who fixated within each ROI at least once was calculated for each visualization. The average proportion of participants who fixated on an ROI (when present) across all of the visualizations is shown in Table 2. In addition, we calculated the proportion of each participant's total fixations that fell within each ROI for each visualization. The average proportion of fixations in each ROI is also shown in Table 2. As before, the three ROIs receiving the highest proportion of fixations were the Data (37%), Title (22%) and Data Label (19%) ROIs.

[6] Due to a programming error, 11 of these images were dropped (leaving a total of 97 images in this experiment). Because they were still of interest, the dropped images were included in a subsequent data collection. The participants in that data collection were recruited in the same manner as the initial group of participants. The group consisted of thirty participants (7 males; mean age = 29.57, stdev = 13.79). Two participants completed both data collection sessions.

The ROIs were categorized based on whether or not they contained text for each stimulus. For each visualization, we then calculated the proportion of fixations that fell in ROIs containing text, the proportion of fixations to the data and data area, and the proportion of fixations that fell in other ROIs that did not contain text (including graphics, symbols, numbers, etc.). On average across all of the visualizations, 40.8% (SD = 19.5%) of the participants' fixations fell into ROIs containing text relative to 44.4% (SD = 18.3%) of fixations in the data ROIs and 14.8% (SD = 0.07%) of fixations in the other non-text ROIs.

Table 2. Attention to each ROI for the visualizations in the second analysis, including average proportions and (standard deviations).

ROI Name	Number of visualizations containing ROI	Average proportion of participants viewing ROI	Average proportion of fixations to ROI
Title	43	0.71 (0.21)	0.22 (0.14)
Data	62	0.91 (0.12)	0.37 (0.18)
Data Area	43	0.53 (0.23)	0.10 (0.06)
X-Axis	46	0.43 (0.18)	0.07 (0.04)
X-Axis Label	23	0.17 (0.11)	0.02 (0.02)
Y-Axis	47	0.52 (0.22)	0.10 (0.10)
Y-Axis Label	33	0.39 (0.23)	0.07 (0.07)
Legend	42	0.68 (0.21)	0.14 (0.08)
Data Label	17	0.70 (0.30)	0.19 (0.13)
Text	24	0.24 (0.29)	0.05 (0.08)

We assessed how often each ROI was one of the first three ROIs fixated by a participant using the same procedure defined above. In this experiment, the Data ROI was most often one of the first three ROIs fixated. It was one of the first three ROIs fixated for 80.5% of the scan paths. The Title ROI was second at 67.5%. Once again, the proportions were lower for the other ROIs (50.8% for Data Labels; 40.5% for Legend; 40.3% for the combination of Y-Axis and Y-Axis Labels; 18.7% for the combination of X-Axis and X-Axis Labels; 13.8% for Text). The axis ROIs were subdivided into those that contained text (other than the axis labels) and those that did not. When the X-Axis ROI contained text, it was one of the first three ROIs viewed in 22.2% of the scan paths. When the X-Axis ROI did not contain text, it was one of the first three ROIs viewed in 14.4% of the scan paths. The Y-Axis ROI was one of the first three ROIs viewed in 56.4% of the scan paths when the ROI included text and 22.0% of the scan paths when it did not.

As before, we also assessed the correlations between the number of words in an ROI and the proportion of fixations in the ROI. The correlations were significant for the Title ($R^2 = 0.90$, $p < 0.001$), Text ($R^2 = 0.81$, $p < 0.001$), X-Axis Label ($R^2 = 0.57$, $p < 0.01$), Y-Axis Label ($R^2 = 0.64$, $p < 0.001$), Legend ($R^2 = 0.39$, $p < 0.02$) and Data Label ($R^2 = 0.60$, $p < 0.02$) ROIs.

As in the first analysis, some of the X- and Y-Axis ROIs contained words and others contained only numbers. For the X-Axis, there was not a significant correlation between the number of items and the proportion of fixations for axes consisting of words ($R^2 = 0.27$, $p = 0.07$) or numbers ($R^2 = 0.03$, $p = 0.86$). For the Y-Axis, there was a significant correlation between the proportion of fixations and the number of words ($R^2 = 0.89$, $p < 0.001$), and, as in the first analysis, a significant negative correlation for numbers ($R^2 = -0.41$, $p < 0.01$).

For a more detailed assessment of how participants allocated their attention to the ROIs, plots were created to show the time course of attention to various parts of the visualizations. Every trial was divided into 313 consecutive 16 ms time windows, from trial onset until the five second trial cutoff time. For each time window, we calculated whether a fixation was made, and if so, which ROI the fixation fell into. An ROI was given a value of 1 for the time window if it received a fixation, and a 0 if it did not. Time windows of 16 ms were chosen to coincide with the sampling rate of the eye-tracker. Fixations were counted as occurring within a time bin if any part of the fixation fell in the window (i.e., even if the fixation ended or started during the time window). Only one fixation was allowed to occur in a single 16 ms time window; if multiple fixations occurred during a time window, only the first ROI visited was counted, and the fixation to the second ROI was assigned as starting in the next time window. However, given that it takes roughly 30-50ms to make a saccade, it is highly unlikely that two separate fixations would have been possible in the small time window. The first fixation of the trial was excluded, as it began with the disappearance of the fixation cross and did not represent a volitional look to any ROI.

The data plotted in Figure 1 shows the viewing patterns collapsing across all visualizations. The x-axis represents time from trial start, the y-axis represents the probability of fixating an ROI, and each line represents a different ROI. Note that the probabilities do not necessarily sum to 1 at every time point, because not every participant made a fixation during every time point (e.g., due to saccades or track loss). Overall, participants tended to look at the Title ROI early in the trial, with Title fixations peaking between 750-1000 ms after trial onset and then quickly declining. Fixations to the Data ROI surpassed looks to the Title beginning ~1500 ms after trial onset, and continued to increase throughout the duration of the trial until peaking at ~4500 ms. The next most-fixated ROI was the Legend region, which had a numerically higher probability of fixation than the rest of the ROIs from ~750 ms after trial onset until the end of the trial. However, the low probability of fixating the other ROIs could be due the fact that not all ROIs were present in all visualizations, meaning that many ROIs had zeros for several visualizations. This plot highlights that although users made more fixations to the data ROI *overall*, this pattern was only true in the later part of the viewing period. Upon first viewing a new visualization, users tended to look at the Title first, after which they shifted their attention to other areas of the visualization.

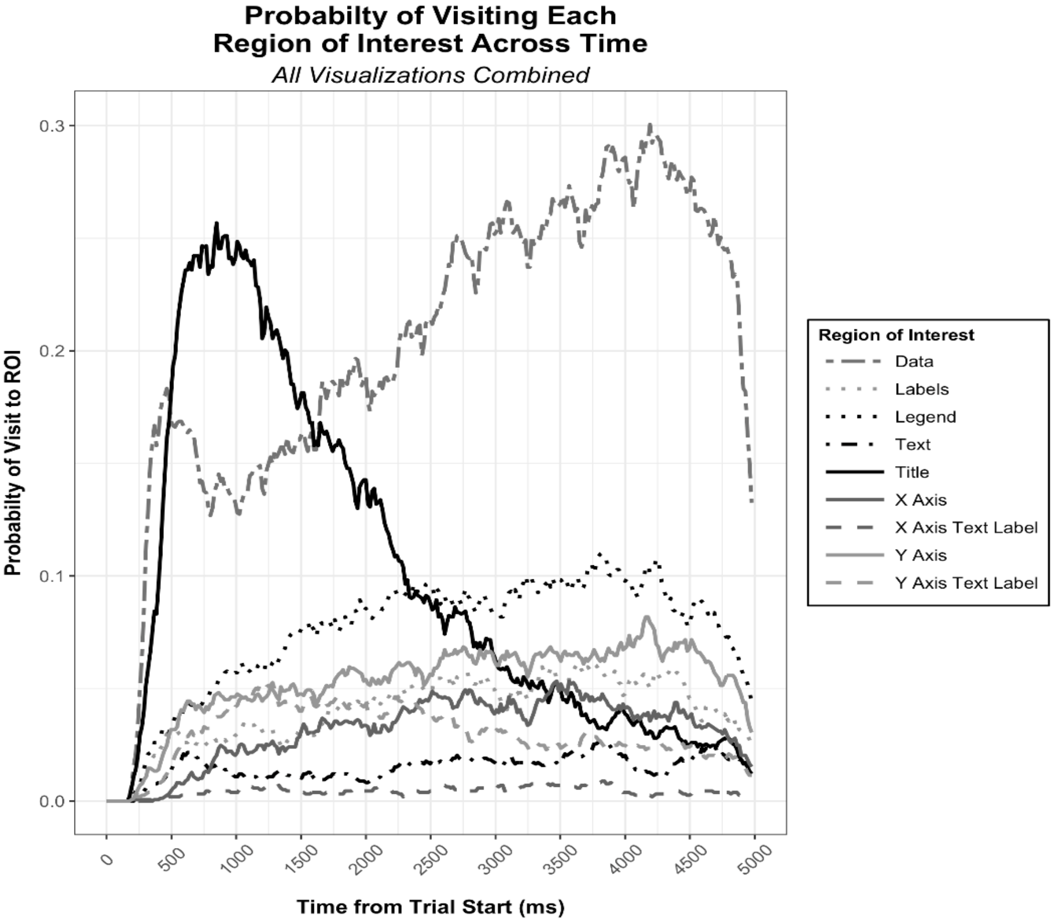


Fig. 1. Probability of fixating each ROI across time, collapsing across all visualizations.

The data plotted in Figure 2 shows viewing patterns to visualizations without text in the y-axis (top panel) versus with text in the y-axis (bottom panel). In both cases, Title fixations peaked early in the trial (~500 ms in vis without y-axis text and ~1000 ms in vis with y-axis text).

However, striking differences are apparent in the pattern of looks to the y-axis. In visualizations *with* y-axis text, users showed clear preference for fixating the y-axis over the data area after ~500 ms into the trial, and fixations to the y-axis exceeded Title fixations after ~2250 ms. Conversely, in visualizations without y-axis text, participants made very few looks to the y-axis, and instead focused most of their fixations on the Title early in the trial, and to the Data ROI later in the trial (after ~1500 ms). There was a small preference for fixating the Labels ROI, relative to the non-Data ROIs, from ~3000-4500 ms, suggesting the need to seek out text to understand the plots when it was not present in the y-axis. This pattern clearly shows that users' viewing patterns to the y-axis were strongly influenced by the presence of text. Users made many more y-axis fixations when text was present compared to when it was not, and even made more fixations to the y-axis than to the Data when text was present, highlighting the emphasis that users place on text during visualization comprehension.

General Discussion

Overall, the results of these analyses suggest that viewers devote a great deal of attention to the text in data visualizations. For the eye tracking data collected as part of the MASSVIS study, the majority of the participants' fixations were devoted to ROIs that contained text. In the second eye tracking dataset, collected using a larger set of data visualizations and a larger group of participants along with a free view rather than memory task, the proportion of fixations devoted to text was comparable to the proportion of fixations devoted to the data.

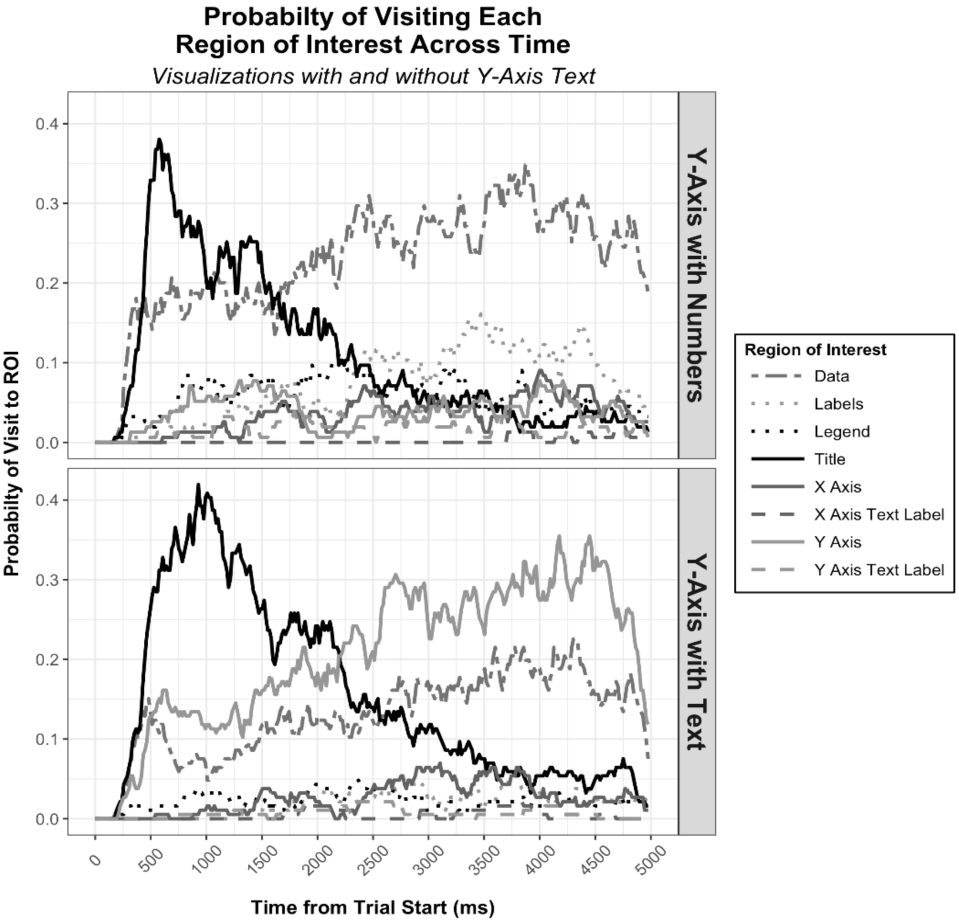


Fig. 2. Probability of fixating each ROI across time, plotted separately for visualizations without y-axis text (top panel) and with y-axis text (bottom panel).

For both datasets, it was instructive to examine the participants' attention to the axes, which contained text in some visualizations and numbers in others. The axes were one of the first three ROIs fixated more often when they contained text than when they did not. Interestingly, for the Y-Axis ROI in both datasets, there was a significant correlation between the proportion of fixations and the number of words in the ROI, and a significant negative correlation between the proportion of fixations and the number of numerical values. An analysis of the time course of fixations for the second dataset indicated that when the Y-Axis ROI contained text, it had a high probability of being visited throughout the trials, and was the most likely ROI to be viewed in the second half of the trials, after participants had turned their attention away from the title of the visualization. When the Y-Axis ROI did not contain text, it had a low probability of visits throughout the trial, with participants devoting more attention to the Data and Legend ROIs.

It is important to note that the two datasets are different in several ways. The MASSVIS data was collected in the context of a memory study where the visualizations were displayed for 10 seconds each. It consisted of visualizations that were found "in the wild." Although we selected a subset of the visualizations that represented common types of data visualizations, these images often contained descriptive titles, annotations, and text noting the source of the data. In other words, the data itself was contextualized by the text in the visualizations. In the second study, we added an additional set of visualizations that were generated in the lab rather than being found in the wild. These visualizations tended to be simpler and had less contextual information. In addition, to mirror the experimental parameters that have been used for assessing visual saliency maps, participants were given a free viewing task[7] with only 5 seconds for examining the visualizations. The simpler text and shorter viewing times in the second dataset may have driven the difference in the overall proportions of fixations to the text versus the data. However, even in the second dataset, the ROIs containing text were viewed almost as often as the data ROIs, indicating that the text still draws viewers' attention even when they have little time and the text provides relatively little information.

Our finding that viewers focused on the text elements in data visualizations is consistent with prior research. Some studies have found that users spend as much as 60-70% of viewing time reading the title, data labels and axes of simple graphs [1, 8, 18]. Users are also more likely to re-fixate text-based areas, such as the legend [3, 22, 29]. In our current analysis, we investigated a wider variety of visualization types and complexities, but the overall tendency to devote a large amount of viewing time to text-based regions remained the same.

The analyses presented here have several limitations. First, the relatively small size of the text in visualizations may necessitate more direct fixations due to the limits of visual acuity [24]. This may have an impact on overall viewing time. Second, the participants in these studies had no particular expertise with interpreting data visualizations, and their tasks did not require them to find specific information in the visualizations, or even to understand the gist of the data presented. While this approach may be realistic for understanding how people process visualizations that they encounter in daily life, such as an infographic presented in a magazine, patterns of attention are likely to be quite different in cases where a viewer is using a visualization to obtain specific information in the context of a larger task. Domain experience also plays an important role in how people attend to data visualizations. Our own prior work found large differences between professional imagery analysts and novice viewers looking at radar imagery [21], and other researchers have found that even brief instructions on how to interpret a plot can change how people allocate their attention [7]. Individual differences in information processing also play an important role. For example, dyslexic individuals spend disproportionately more time on text than typical readers [18]. None of these factors operate in isolation, and taking their combination into account can result in complex interactions between such factors as chart type, task difficulty, and the user's perceptual speed [29].

Despite these limitations, the general finding that text in data visualizations draws the viewer's attention has important implications for the development of visual saliency models that apply to visualizations. As discussed above, the ability to make predictions about where viewers will look in data visualizations could be a useful evaluation tool. To make accurate predictions, these models must take attention to text into account. In our future work, we plan to develop a new saliency model that incorporates text as a visual feature. We will test how to weight this feature relative to the other visual features that are commonly used in saliency models (color, contrast, and orientation). If successful, this approach will provide an improved tool that will allow visualization designers to evaluate their designs from the perspective of human visual processing.

Acknowledgements.

This work was funded by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

The authors would like to thank Deborah Cronin and Jim Crowell for collecting the eye tracking data at the University of Illinois at Urbana-Champaign, as well as Hank Kaczmarski and Camille Goudeseune for their support.

References

Acarturk, C., Habel, C., Cagiltay, K., & Alacam, O.: Multi-media comprehension of language and graphics. J Eye Mov Res **1**(3):2, 1-15. (2008). doi: 10.16910/jemr.1.3.2

Borji, A., & Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. CVPR 2015 workshop on "Future of Datasets" (2015), arXiv preprint arXiv:1505.03581

Borkin, M., Bylinskii, Z., Kim, N., C.M., B., Yeh, C., Borkin, D., Pfister, H., & Oliva, A.: Beyond memorability: Visualization recognition and recall. IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis) (2015). doi: 10.1109/TVCG.2015.2467732

Borkin, M., Vo, A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. What makes a visualization memorable? In IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis) (2013). doi: 10.1109/TVCG.2013.234

Bylinskii, Z., & Borkin, M. A.: Eye fixation metrics for large scale analysis of information visualizations. In Proceedings of ETVIS 2015, First Workshop on Eyetracking and Visualizations (2015).

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A.: MIT saliency benchmark. http://saliency.mit.edu/

[7] It is worth noting that a free viewing task may be more representative of how people interact with visualizations in the wild than a memory task. When a person encounters a data visualization in *The Economist*, for example, they are essentially doing a free viewing task.

Canham, M., & Hegarty, M.: Effects of knowledge and display design on comprehension of complex graphics. Learn Instr **20**, 155-166. (2010). doi: 10.1016/j.learninstruc.2009.02.014

Carpenter, P. A., & Shah, P.: A model of the perceptual and conceptual processes in graph comprehension. J Exp Psychol Appl **4**(2), 75–100. (1998). doi: 10.1037//1076-898x.4.2.75

Connor, C.E., Egeth, H.E., & Yantis, S.: Visual attention: Bottom-up versus top- down. Curr Biol **14**(19), R850–R852. (2004). doi: 10.1016/j.cub.2004.09.041

Fu, B., Noy, N. F., & Storey, M. A.: Eye tracking the user experience–An evaluation of ontology visualization techniques. Semant Web **8**(1) 23-41. (2017). doi: 10.3233/SW-140163

Goldberg, J. H., & Helfman, J. I.: Comparing information graphics: A critical look at eye tracking. In Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization, pp. 71-78. (2010). doi: 10.1145/2110192.2110203

Goldberg. J. H., & Helfman, J. I.: Eye tracking for visualization evaluation: Reading values on linear versus radial graphs. Inf Vis **10**(3), 182-195. (2011). doi: 10.1177/1473871611406623

Haass, M. J., Wilson, A. T., Matzen, L. E., & Divis, K. M.: Modeling Human Comprehension of Data Visualizations. In International Conference on Virtual, Augmented and Mixed Reality, pp. 125-134. (2016). doi: 10.1007/978-3-319-39907-2_12

Higgins, E., Leigenger, M., & Rayner, K.: Eye movements when viewing advertisements. Front Psychol, **5**, 210. (2014). doi: doi.org/10.3389/fpsyg.2014.00210

Ishihara, S. (1972). Tests for Colour-Blindness; 24 Plates Edition. Tokyo: Kanehara Shuppan Co., Ltd.

Itti, L., & Koch, C.: Computational modelling of visual attention. Nat Rev Neurosci **2**, 194-203. (2001). doi: 10.1038/35058500

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. [https://dspace.mit.edu/handle/1721.1/68590]

Kim. S., & Lombardino, L. J.: Comparing graphs and text: Effects of complexity and task. J Eye Mov Res **8**(3):2, 1-17. (2015). doi: 10.16910/jemr.8.3.2

Kim. S., Lombardino, L. J., Cowles, W., & Altmann, L. J.: Investigating graph comprehension in students with dyslexia: An eye-tracking study. Res Dev Disabil, **35**, 1609-1622. (2015). doi: 10.1016/j.ridd.2014.03.043

MATLAB Release 2015b, The MathWorks, Inc., Natick, Massachusetts, United States

Matzen, L. E., Haass, M. J., Tran, J. & McNamara, L. A.: Using eye tracking metrics and visual saliency maps to assess image utility. Electronic Imaging **16**, 1-8. (2016). doi: doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-127

Peebles, D., & Cheng. P. C.-H.: Modeling the effect of task and graphical representation on response latency in a graph reading task. Hum Factors **45**(1), 28-46. (2003). doi: 10.1518/hfes.45.1.28.27225

Pinto, Y., van der Leij, A., Sligte, I. G., Lamme, V. A. F., & Scholte, H. S.: Bottom-up and top-down attention are independent. J Vis **13**, 1-14 (2013). doi: 10.1167/13.3.16

Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol Bull **124**(3), 372-422. (1998). doi: 10.1037/0033-2909.124.3.372

Rayner, K.: Eye movements and attention in reading, scene perception, and visual search. Q J Exp Psychol **62**(8), 1457-1506. (2009). doi: 10.1080/17470210902816461

Rosenholtz, R., Dorai, A., & Freeman, R.: Do predictions of visual perception aid design? In ACM Transactions on Applied Perception (TAP), **8**(2), 12. (2011). doi: 10.1145/1870076.1870080

Shah, P., & Hoeffner, J.: Review of graph comprehension research: implications for instruction. Educ Psychol Rev **14**(1), 47-69. (2002). doi: 10.1023/A:1013180410169

Strobel, B., Sass, S., Lindner, M. A., & Köller, O.: Do graph readers prefer the graph type most suited to a given task? Insights from eye tracking. J Eye Mov Res **9**(4):4, 1-15. (2016). doi: 10.16910/jemr.9.4.4

Toker, D., Conati, C., Steichen, B., & Carenini, G.: Individual user characteristics and information visualization: connecting the dots through eye tracking. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 295-304. (2013). doi: 10.1145/2470654.2470696

APPENDIX E: INFLUENCES OF TASK GOALS AND EXPERIENCE ON VIEWING ABSTRACT DATA VISUALIZATIONS

Kristin M. Divis[1], Laura E. Matzen[1], Deborah A. Cronin[2], Michael J. Haass[1]

[1]Sandia National Laboratories, [2]University of Illinois at Urbana-Champaign

Introduction

Data visualizations serve an important role in scientific inquiry and communication. A good data visualization can allow its viewer to quickly identify important trends and interesting groups or outliers in a large dataset or to rapidly grasp the take-home message of an entire study. But what makes a data visualization "good"? Members of the visualization community are calling for evaluation of visualizations by examining the extent to which they support their viewers' cognitive needs (Cleveland, 1993; Card, Mackinlay, and Shneiderman, 1999; Etemadpour, Olk, & Linsen, 2014; Gleicher, et al., 2013; Micallef, et al., 2017; Munzner, 2014). Under this form of metric, a "good" visualization successfully exploits its viewers' cognitive processes to draw the users' attention to relevant information, minimize distraction, and increase the likelihood of correct interpretation.

One method for evaluating whether a visualization supports its users' cognitive processes and ensures the users are utilizing the visualization as the visualizer intended is to examine the users' eye movements while they view the visualization. The structure of the retina limits visual acuity to the central portion of our field of view. In order to extract detailed information about a region in space we are not currently looking at, we must move our eyes to that region. Thus, when users move their eyes to a particular area on a data visualization, we can trust they are attending that region (Hoffman & Subramaniam, 1995) and that they are extracting and processing information that is available within it (Aloimonos, Weiss, & Bandyopadhyay, 1987; Findlay & Gilchrist, 2003; Henderson, 2003). The pattern of eye movements and the pauses between each eye movement (fixations) can inform our understanding of how a user is processing a visualization. Furthermore, if we understand the factors guiding attention, we can develop models to evaluate whether the most relevant portions of the visualization are likely to draw a hypothetical user's attention without undergoing expensive and time-consuming user studies.

Recently, several research groups have suggested visual salience may be a useful tool for evaluating the extent to which visualizations support their users' cognitive processes (Janicke & Chen, 2010; Kim & Varshney, 2006; Matzen et al., 2017). Human visual attention is drawn automatically to unique (salient) features in the visual array. Models of human attention based on visual salience predict participants' eye movements with good success in natural (e.g., Itti & Koch, 2001) and man-made (e.g., Berg & Itti, 2008) scenes under free-viewing conditions. As such, optimal utilization of visual salience in a data visualization may guide users through the visualization in the way the visualizer intended while sub-optimal usage may distract the user. Indeed, increasing the visual salience of task-relevant information has been shown to alter or aid user performance in day-to-day decision-making tasks (Milosavljevic et al., 2012), in virtual reality (Veas et al., 2011), and, importantly, in data visualization interpretation (Gleicher et al., 2013; Healey & Enns, 1998; Hegarty, et al., 2010; Interrante, 2000; Nothelfer, Gleicher, & Franconeri, 2017). A modified version of Itti & Koch's (2001) model of visual salience has also been used to successfully predict users' eye movements in data visualizations (Matzen et al., 2017).

Visual salience is one of several external sources of information that guide human attention in a "bottom-up" fashion—highly salient areas pull attention without high-level cognitive input by the observer. Attention can also be guided in a "top-down" fashion, whereby the observer's goals, expectations, and prior experience guide attention to goal-relevant objects (Yarbus, 1967; Wolfe, 1994;

Wolfe, Cave, & Franzel, 1989). Users viewing a simple scatterplot, therefore, may have their attention pulled automatically towards a cluster of salient, red data points amongst several grey clusters or may deploy their attention voluntarily to the low-salience axes labels because their prior experiences and knowledge suggests they should.

While visual salience can predict where users of data visualization will look under some circumstances (Matzen et al., 2017), top-down processing has a strong influence over where people choose to look and can override the bottom-up draw of visually salient regions (Land & Hayhoe, 2001; Land, Mennie, & Rusted, 1999). For instance, social cues (Birmingham, Bischof, & Kingstone, 2009), differing task priorities (Castelhano, Mack, & Henderson, 2009; Foulsham & Underwood, 2007; Hegarty et al., 2010; Henderson, et al., 2007; Land, Mennie, & Rusted, 1999; Mills et al., 2011), expertise (Lansdale, Underwood, & Davies, 2010), and prior experiences (Chun, 2000; Lleras, Rensink, & Enns, 2005, 2008) are all well-known top-down drivers of attention. Recent research has also demonstrated the influence of top down goals on eye movement patterns when interpreting data visualizations. Participants with different goals looking at the same visualization will inspect it differently (Michal & Franconeri, 2017; Michal, Uttal, Shah, & Franconeri, 2016).

To better evaluate the extent to which a visualization supports a user's cognitive processes, the factors that guide top-down attention should be considered in addition to bottom-up factors like visual salience. As an example, Matzen and colleagues (2017) developed the Data Visualization Saliency (DVS) model of visual salience for evaluating data visualizations that took into account the importance of text. Text typically has low visual salience and is often not identified by salience-based models of human attention (e.g., Itti & Koch, 2001) as a region of interest. However, humans have a top-down motivation to look at text despite its low visual salience, and most if not all viewers will choose to look at text if it is present (Rayner et al., 2001), particularly in the context of interpreting data visualizations (Matzen et al., 2016). Thus, the DVS model better predicted participants' eye movements for data visualizations than other models that did not take into account any top-down processes. There are likely other common components of data visualizations that users consistently attend in a top-down fashion, as they do with text. If this is the case, identifying these visualization components and adding sensitivity to those components to a salience-based model should further improve the model's performance.

Creating a useful evaluation of this sort for all data visualizations is difficult in part because data visualizations are developed and consumed for a wide variety of high-level goals. For example, Munzner (2014) suggests data visualizations can be created with the broad goals of analyzing, searching, or querying information. Within each of those broad goals lies a number of more specific goals—a user analyzing data with a data visualization may be seeking to discover new information, a user searching data via visualization may be exploring information to locate targets of interest, or a user querying via data visualization may be seeking to compare two data sets. Each of these goals and combinations of them is best served by different kinds of data visualizations and, even within the same data visualization, users with different goals may choose to look at different things. Characterizing the eye movement behavior associated with certain user goals could further inform a model for evaluating visualizations. For example, if the purpose of a visualization is to allow the user to compare two pieces of information, a model that accounts for that goal could adjust the weighting of certain features to create an output that better reflects that top-down driven goal.

Toward this end, the goal of the present study is to better understand the role of top-down attention in comprehension of data visualizations. If two users have different goals when viewing the same data visualization, how do their behaviors differ and what behaviors are similar? Mills and colleagues (2010) compared eye movement behavior for four different tasks in natural scenes. When they compared the fixation durations and eye movement amplitudes of their participants under different task conditions, they found characteristic patterns of eye movement behavior for each task type. The eye-movement patterns for these tasks were similar in some ways (e.g., a visual search task and an aesthetic evaluation task produced similarly short fixation durations) and different in others (e.g., the visual search task was better characterized by longer eye-movements than the aesthetic evaluation task). Applying a similar method to data visualizations will provide a better understanding of how top-down factors guide eye movements through a visualization, thus opening the door to visualization evaluation techniques informed by the user's top-down goals.

Here, we describe two studies in which we investigate the impact of top-down factors on data visualization comprehension. In both experiments, participants' eyes were tracked. In Experiment 1A, we presented participants with scatterplots and asked them to describe the trend they observed or to describe any outliers present in the display. Each participant was presented with the same stimuli regardless of task instruction, so we were able to directly compare eye movement patterns and behavioral results for the two tasks for each scatterplot. In Experiment 1B, participants were given two clusters of data and were asked to judge the membership of an intermediate reference point (i.e., which cluster does the point belong to) or to judge which cluster was higher. The choice of tasks for Experiments 1A and 1B were informed by two prior studies investigating the ways perceptual information influences how users interpret scatterplots. Etemadpour, Olk, & Linsen (2014) evaluated how users' eye movement patterns and behavioral responses changed as a function of the layout and design of plots. Gleicher and colleagues (2013) investigated the influence of group size, group encoding style (including the salience of the group's encoding), and the number of groups on participants' ability to make mean-value judgements for one group within a multi-class scatterplot. They found that numerosity of group members, the number of total groups, and redundant encodings (i.e., one group is demarked two unique encodings) did not affect participants' ability to make mean value judgements. However, low salience encodings of group membership negatively impacted participants' performance. Experiments 1A and 1B extend these findings.

In Experiment 2, we sought to evaluate the extent the top-down factor of expertise influences users' eye movements and behavioral performance when viewing data visualizations. We recruited participants experienced with statistics and presented them with a variety of styles of data visualizations (e.g., scatterplot, box and whisker plot, violin plot, etc.) and a variety of representations of statistical uncertainty (e.g., standard error of the mean, interquartile range, etc). Each stimulus depicted two groups. Participants evaluated whether they felt the two groups presented in each graph were statistically significantly different and then rated their certainty in their evaluation. The same data was depicted using each of the chart types and representations of uncertainty, allowing us to examine the extent to which participants were able to evaluate the statistical significance of the difference between the two groups given a certain data visualization (i.e., the visualizations' efficacy compared to other visualizations of the same data). We also asked participants to indicate their familiarity with each of the chart and representations of uncertainty used in the experiment to further evaluate the influence of expertise on participants' eye movements and behavioral performance.

Experiment 1A

In Experiment 1A, we examined participants' eye movements and behavioral responses while viewing simple scatterplots with the goal of either detecting outliers or describing the relationship between the two variables plotted.

Methods

Participants

Thirty participants were recruited from students, faculty, and staff in the University of Illinois community (7 males; *mean age* = 29.57, *stdev* = 13.79) and compensated $20 for their time. All participants were tested for color vision deficiencies (24 plate Ishihara Test; Ishihara, 1972) and near vision acuity prior to completing the study.

Design
Task (trend or outlier description), data pattern (+/- linear, sinusoidal, +/- logarithmic, flat, and +/- quadratic), and number of outliers (2 or 4) were manipulated within subjects.

Materials
All stimuli were created in R Software (R Development Core Team, 2008) from simulated data, using the standard plotting function to create simple scatterplots. The stimuli were plotted on a white background, with labeled axes and main title. Each stimulus consisted of 100 data points (plotted as open circles). All foreground elements were black in color. See Appendix A for example stimuli.

Thirty-two scatterplots with the following trends were created for Experiment 1A: positive linear, negative linear, flat, sinusoidal (cyclical), positive logarithmic (asymptotic), negative logarithmic (asymptotic), positive quadratic, and negative quadratic. Each graph had either 2 or 4 outliers. The 32 stimuli consisted of 2 each of the 16 unique combinations of trend and number of outliers. Simulated data were drawn from Gaussian distributions with intuitive parameters for the given axis labels. The main body of data was constrained to fall within two vertical standard deviations of the trend function. The outliers were created to be at least four standard deviations away from the trend function.

Procedure

The experiment was completed individually in a dark room at a nominal viewing distance of .8 meters. Stimuli were presented on a large monitor (.932 x .523 meters; 1920 x 1080 pixels) while eye movements were recorded with two Smart Eye Pro cameras. Participants first underwent the standard Smart Eye camera setup procedure and 9-point calibration.

Experiment 1A was divided into 3 sections: practice and two blocks of stimuli. During the practice session, participants worked through two example stimuli and were given the opportunity to ask the experimenter for further clarification. Half the participants described the data trend during the first block of stimuli and then described the outliers during the second; the other half of the participants described the outliers in the first block and then described the data trend in the second block.

The 32 images were divided into two sets, each containing 16 unique combinations of data trends and number of outliers (counterbalanced across the two blocks of stimuli between subjects). Each image was 1000 pixels high (width was allowed to vary to maintain aspect ratio) and placed in the center of the screen; the edges were white-padded to fill the screen.

Each image was presented one at a time, was preceded with a fixation cross, and had a 500 ms interstimulus interval. Participants were allowed to work through the images at their own pace, with a maximum of 10 seconds allowed on each image. After studying the image, participants advanced to a blank screen and verbally described the image (either the trend or the outliers, depending on condition). The experimenter recorded their response and asked for further clarification if necessary (all responses were also captured in audio files).

Behavioral Results

All statistical tests reported here were held at an $\alpha = .05$ level and run using R Software (R Development Core Team, 2008).

Trend Description

This section covers the blocks where participants described the trend shown in the scatterplot. Each response was scored on a scale from *1* to *3*, with *1* meaning the response did not demonstrate sufficient understanding of the trend, *2* meaning the participants noted some of the key features but were partially incorrect or did not describe it in full, and *3* meaning the response demonstrated an acceptable level of understanding. Because this scoring system is subjective, we had two raters independently score each response. On responses where there was a discrepancy in the scores, the raters discussed their scores. If an agreement was reached, the score was updated; if an agreement was not reached, the scores were not changed. After this process, the correlation between the two sets of scores was quite high ($r = .958$). Subsequent analyses used the average of the two scores. Subsequent analyses used the average of the two scores.

A Kruskal-Wallis rank sum test indicated a significant difference in scores among the types of graphs ($X^2 = 97.752$, $df = 4$, $p < .001$).

Outlier Detection

This section covers the blocks where participants identified outliers in each image. Once again, two raters independently worked through each response, indicating how many outliers were identified. The counts from the two raters were compared; where the counts differed, the raters discussed the response and updated their score if they came to an agreement. Ambiguous responses or those that did not mention outliers were flagged and dropped from subsequent analyses. If one rater indicated that the response was too ambiguous to score, it was also dropped from subsequent analyses. The correlation between the two raters was once again quite high ($r = .961$). Errors were calculated as the difference between the actual number of outliers (2 or 4) and the number of outliers reported (average of the counts from the two raters).

Out of 398 trials, errors were made on 220 trials (55.3%). Participants tended to miss outliers (200 trials, 90.9% of errors) rather than falsely identify outliers. A Kruskal-Wallis rank sum test indicated no significant difference in absolute error among the types of graphs ($X^2 = 7.697$, $df = 4$, $p = .103$).

Eye Movement Results

Fixations were calculated using SmartEye's default algorithm (any sample for which the velocity over the preceding 200 ms is less than 15°/s is deemed a fixation). Any fixation less than 100 ms and first fixations in each trial were dropped.

In Experiment 1A, our primary interest was in how visual attention changes in response to differences in task. A mixed effects model with a fixed effect for task and random intercepts for participant and stimulus (using Satterthwaite approximation for degrees of freedom) revealed that overall, participants had more fixations in the outlier task (*mean* = 22.83 fixations, *stdev* = 4.78) relative to the trend task (*mean* = 19.90 fixations, *stdev* = 5.37; $t(885) = 10.04$, $p < .001$). A similar mixed effects model with fixation duration as the fixed effect revealed that fixation durations in the trend task (*mean* = 325.38 ms, *stdev* = 293.92) tended to be longer than those in the outlier task (*mean* = 279.84 ms; *stdev* = 232.53; $t(20015) = 12.43$, $p < .001$).

Task also influenced which regions of the graph participants most frequently fixated. Each stimulus was divided into the following regions of interest (ROIs): outliers, trend, title, x-axis, x-label, y-axis, y-label, and other. Proportion of fixations to each type of ROI were calculated for each participant and stimulus (see Figure KK). The critical ROIs of interest were the trend and outlier ROIs due to their direct relevance to the two tasks of trend description and outlier detection. A mixed effects model predicting proportion of fixations as a function of the fixed effects of task and ROI and with random intercepts for subject and stimulus (using Satterthwaite approximation for degrees of freedom) revealed significant simple effects of task for both: a higher proportion of fixations occurred to the outlier ROIs in the outlier task ($t(7560) = 4.41$, $p < .001$), but a higher proportion of fixations occurred to the trend ROI in the trend task ($t(7560) = 12.71$, $p < .001$). The model also revealed that participants in the trend description task (relative to the outlier detection task) had a higher proportion of fixations to all other ROIs, with the exception of the "other" ROI where those in the outlier detection task had a higher proportion of fixations (all t-statistics < 2.00 and p-values < .05).

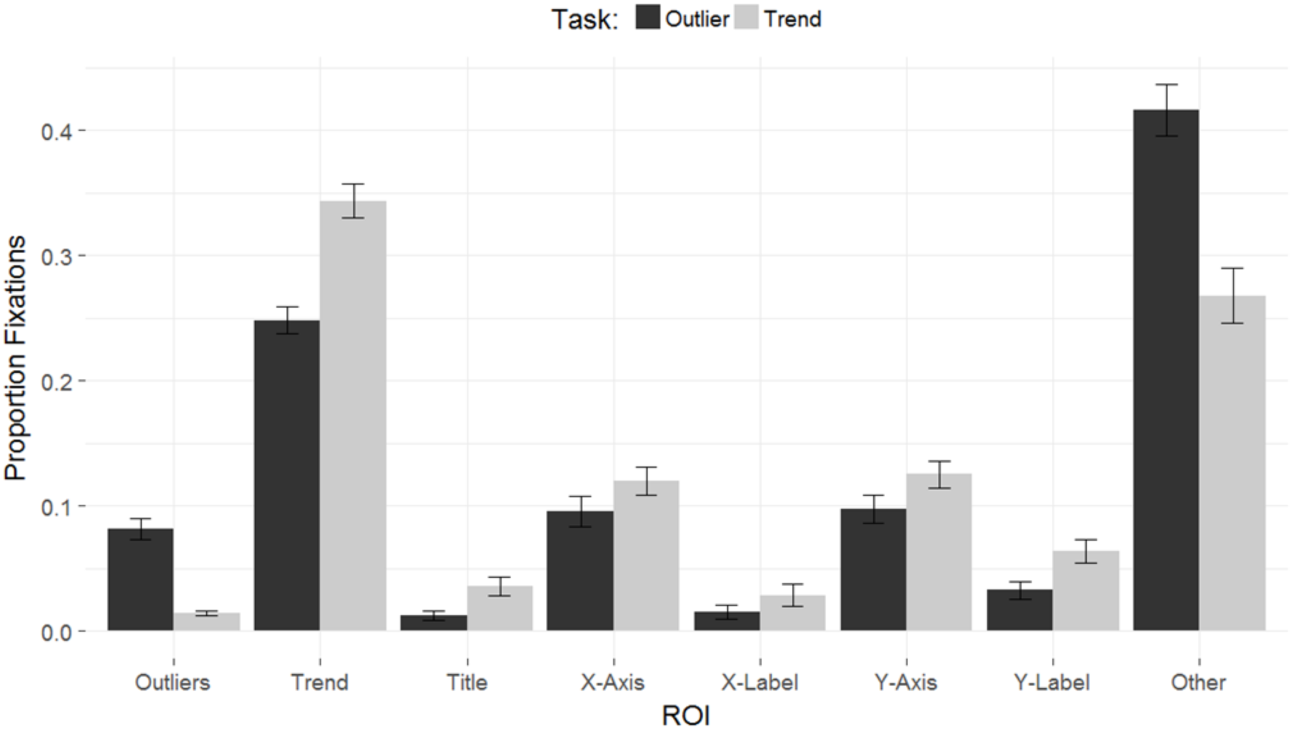


Figure KK. Proportion of fixations each subject made to regions of interest (ROIs) based on task (outlier detection or trend description) in Experiment 1A. Error bars represent standard error of the mean.

Experiment 1B

In Experiment 1B, we monitored participants' eye movements while viewing two-dimensional clusters of data with a reference point superimposed between the clusters. Participants were asked to either identify which cluster the reference point belongs to or which cluster has the overall highest vertical mean.

Methods

Participants

The same participants who completed Experiment 1A also completed Experiment 1B.

Design

Task (cluster mean comparison or reference point grouping), reference point centering (standard deviation or mean), relative cluster height (even or raised), cluster sparsity (low or high), and cluster dispersion (low or high) were manipulated within subjects.

Materials

All stimuli were created in R Software (R Development Core Team, 2008) from simulated data using the *ggplot2* software package (Wickham, 2009). The scatterplots had design characteristics similar to those used in Etemadpour, Olk, and Linsen (2014). Each scatterplot had two clusters and one reference point. The data points were filled colored circles outlined in black on a white background. No axis titles or tick marks were provided. One cluster was blue and the other was green (randomly assigned); the reference point was always red. See Appendix B for example stimuli.

Clusters were manipulated along the sparsity (low or high) and dispersion (low or high) dimensions. Clusters with high sparsity contained fewer data points per square unit than those with low sparsity. Clusters with high dispersion were more spread out (i.e., higher standard deviation, leading to a wider cluster) than those with low dispersion. Crossing these two dimensions leads to four types of clusters: *Cluster A*: low sparsity and low dispersion (n = 40, *stdev* = 10), *Cluster B*: high sparsity and low dispersion (n = 15, *stdev* = 10), *Cluster C*: low sparsity and high dispersion (n = 85, *stdev* = 25), and *Cluster D*: high sparsity and high dispersion (n = 40, *stdev* = 25). Simulated data were drawn from Gaussian distributions with the parameters indicated in each cluster class. See Appendix B for examples of each cluster type. Clusters were paired in all possible combinations (e.g., A-A, C-D, D-C) to create 80 total images. In half of the stimuli, the mean cluster height was the same for each cluster in the pair; in the other half, one cluster was higher than the other.

A reference point was placed between the two clusters. One cluster was always to the left of the reference point; the other was always to the right. The reference point was mean-centered on 50% of the stimuli and standard-deviation-centered on the other 50%. When the reference point was mean-centered, it was exactly halfway between the horizontal and vertical mean for both clusters. When the reference point was standard-deviation-centered, it was exactly four standard deviations along the horizontal axis away from the mean of each cluster (and mean-centered along the vertical axis for clusters with means at the same height or one vertical standard deviation above or below the means of the clusters for clusters at different heights).

Procedure

Experiment 1B was completed following a short break after Experiment 1A. It used the same setup as in Experiment 1A.

Experiment 1B was also divided into 3 sections: practice and two blocks of stimuli. During the practice session, participants worked through two example stimuli and were given the opportunity to ask the experimenter for further clarification. Half the participants indicated which cluster's mean was higher in the first block of stimuli and then indicated which cluster the reference point belonged to in the second; the other half of the participants indicated reference point membership in the first block of stimuli and then indicated which cluster was higher in the second block.

The 80 images were divided into two sets, each containing 40 images (with cluster pairing, reference point centering, and relative cluster height counterbalanced). The groups were counterbalanced across Sets 1 and 2 and between subjects (see Appendix B). Each image was 1000 pixels high (width was allowed to vary to maintain aspect ratio) and placed in the center of the screen; the edges were white-padded to fill the screen.

Each image was presented one at a time, was preceded with a fixation cross, and had a 500 ms interstimulus interval. Participants were allowed to work through the images at their own pace, with a maximum of 10 seconds allowed on each image. Participants pressed a key to indicate which cluster was higher or which cluster the reference point belonged to (depending on condition). The experiment advanced to the next image after the key press.[8]

Behavioral Results

All statistical tests reported here were held at an $\alpha = .05$ level (95% confidence interval, *CI*). Exact binomial tests analyzed whether the clusters chosen differed significantly from what one would expect based on chance (50%). All analyses were run using R Software (R Development Core Team, 2008).

Reference Point Membership

On half of the trials, participants were asked to indicate which cluster (left or right) the reference point belonged to. The analyses in this subsection are for that reference point membership task.

Across all stimuli, participants showed a slight bias toward indicating the reference point belonged to the cluster on the right (53.6%, *CI* [50.7%, 56.4%], $p = .014$). However, all conditions were perfectly counterbalanced across the left-right dimension, so this bias does not systematically change the interpretation of the results. All further analyses are collapsed across whether the cluster was on the left or right side of the screen.

Sparsity

Selecting trials in which one cluster had low sparsity and one had high sparsity (more vs. fewer data points per square unit), we analyzed whether relative sparsity of the clusters influenced participants' decisions in the reference point membership task.

Overall, participants consistently indicated that the reference point belonged to the cluster with lower sparsity (more data points per square unit). The cluster with lower sparsity was chosen 78.8% of the time (*CI* [73.0%, 83.7%], $p < .001$). This pattern held, regardless of whether the clusters also had low dispersion (low sparsity chosen 81.7%, *CI* [73.6%, 88.1%], $p < .001$) or high dispersion (low sparsity chosen 75.8%, *CI* [67.2%, 83.2%], $p < .001$). It also held regardless of whether the reference point was mean centered (low sparsity chosen 78.3%, *CI* [69.9%, 85.3%], $p < .001$) or standard deviation centered (low sparsity chosen 81.7%, *CI* [70.8%, 86.0%], $p < .001$).

Dispersion

We also examined the influence of low versus high dispersion (how spread out the points were) on participants' preference for reference point cluster membership.

When collapsing across sparsity (low *vs.* high) and centering (mean *vs.* standard deviation), no significant effects were found (high dispersion cluster chosen 52.1%, *CI* [45.6%, 58.6%], $p = .561$). However, that null result appears to have been driven by reference point centering technique leading to opposite effects. When the reference point was mean centered, participants were more likely to indicate the reference point belonged to the cluster with a *high* dispersion (high dispersion cluster chosen 91.7%, *CI* [85.2%, 95.9%], $p < .001$). This pattern held, regardless of whether the clusters had low sparsity (high dispersion cluster chosen 83.3%, *CI* [71.5%, 91.7%], $p < .001$) or high sparsity (high dispersion cluster chosen 100.0%, *CI* [94.0%, 100.0%], $p < .001$). When the reference point was standard deviation centered, participants were more likely to indicate the reference point belonged to the cluster with a *low* dispersion (low dispersion cluster chosen 87.5%, *CI* [80.2%, 92.8%], $p < .001$). Once again, this pattern held regardless of whether the clusters had low sparsity (low dispersion cluster chosen 91.7%, *CI* [81.6%, 97.2%], $p < .001$) or high sparsity (low dispersion cluster chosen 83.3%, *CI* [71.5%, 91.7%], $p < .001$).

Cluster Types

[8] Following a short break after completing Experiment 2, participants also worked through a block a free view data visualization images while their eyes were tracked. They were asked about their experience using graphs (verbal explanation and 5-point Likert rating from very infrequently to very frequently interpret graphs). Those results were not analyzed as part of this manuscript and are therefore not included.

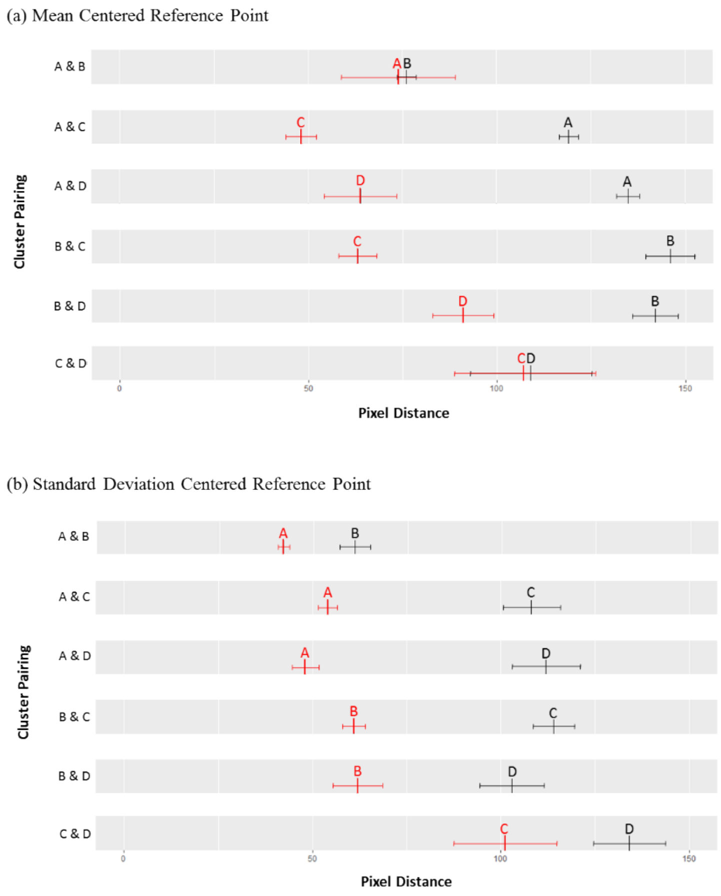
We also analyzed which cluster was preferred for reference point membership when different cluster types were pitted against one another. Recall that clusters of type *A* had low dispersion and low sparsity, *B* had low dispersion and high sparsity, *C* had high dispersion and low sparsity, and *D* had high dispersion and high sparsity. See Table X for the results, broken down by reference point centering method. For stimuli that had mean centered reference points, the preferred cluster type for reference point membership ranked as follows: 1. *C*, 2. *D*, 3. *A*, and 4. *B*. For stimuli that had standard deviation centered reference points, the preferred cluster type for reference point membership ranked as follows: 1. *A*, 2. *B*, 3. *C*, and 4. *D*. The driving factor for preference was dispersion, but its effect differed based on centering method. Clusters with high dispersion were preferred when the reference point was mean centered; clusters with low dispersion were preferred when the reference point was standard deviation centered. Of secondary importance was sparsity, with low sparsity clusters preferred over high sparsity clusters.

Centering	Clusters	Chosen Most	Percent	95% CI	p-value
Mean	A vs. B	A	78.3%	[65.9%, 87.9%]	< .001
	A vs. C	C	83.3%	[71.5%, 91.7%]	< .001
	A vs. D	D	81.7%	[69.6%, 90.5%]	< .001
	B vs. C	C	91.7%	[81.6%, 97.2%]	< .001
	B vs. D	D	100.0%	[94.0%, 100.0%]	< .001
	C vs. D	C	78.3%	[65.9%, 87.9%]	< .001
Standard Deviation	A vs. B	A	85.0%	[73.4%, 92.9%]	< .001
	A vs. C	A	91.7%	[81.6%, 97.2%]	< .001
	A vs. D	A	93.3%	[83.8%, 98.2%]	< .001
	B vs. C	B	85.0%	[73.4%, 92.9%]	< .001
	B vs. D	B	83.3%	[71.5%, 91.7%]	< .001
	C vs. D	C	73.3%	[60.3%, 83.9%]	< .001

Table X. *Cluster types A (low dispersion, low sparsity), B (low dispersion, high sparsity), C (high dispersion, low sparsity), and D (high dispersion, high sparsity), compared in the reference point membership task and split based on reference point centering method. Percentage for most common choice, 95% condifidence intervals, and p-value reported.*

Nearest Neighbor

One explanation for participants' decisions in this task is they pair the reference point with the cluster that has the closest point to that reference point. To investigate, we used GIMP software (GIMP Development Team, 2007) to hand code the pixel coordinates of the center of the reference point and the nearest neighbor point in each cluster for each stimulus. We calculated the distance between the nearest neighbor in each cluster and the reference point. We noted which cluster had the nearest point and determined whether that cluster was chosen by participants more often than the other cluster. For 91.3% of the stimuli, the cluster with the nearest neighbor was chosen more frequently than the other cluster (*CI* [82.8%, 96.4%], *p* <.001). See Figure ZZ for average distance between nearest neighbor and reference point for stimuli with mean centered and standard deviation centered reference points. Notably, the average nearest neighbor metric perfectly aligns with the cluster type preferred. Because the nearest neighbor distance and the distribution/weight of the cluster (based on sparsity and dispersion) are highly correlated, the current study does not lend itself to teasing apart the contribution of each.



(a) Mean Centered Reference Point

(b) Standard Deviation Centered Reference Point

Figure ZZ. Average pixel distance between nearest neighbor and reference point when reference point was (a) mean centered and (b) standard deviation centered. The cluster chosen most often as being closest to the reference point is highlighted in red. Error bars represent standard deviation of the mean.

Discussion

Dispersion, combined with reference point centering method, was the driving factor in the reference point membership task. When the reference point was mean centered, participants chose the cluster with high dispersion (higher standard deviation). When the reference point was standard deviation centered, participants chose the cluster with low dispersion (low standard deviation). Notably, the reference point centering method only mattered when the two clusters were drawn from distributions with different standard deviations. See Figures B2 and B3 in Appendix B for examples of clusters with the similar and different variances. The combination of dispersion and centering method goes hand in hand with nearest neighbor metric. When the reference point was mean centered, the cluster with the closest nearest neighbor tended to be the one with high dispersion; the opposite pattern held for standard deviation centered reference points. Sparsity also had a consistent effect on the task, although it fell to secondary importance after dispersion. Participants generally chose the cluster with low sparsity (more data per square unit) as opposed to high sparsity.

Highest Cluster

On half the trials in Experiment 1B, participants performed the reference point membership task; on the other half, they indicated which of the clusters was higher. In half the stimuli, one of the clusters had a higher (y-axis) mean value; in the other of the stimuli, the vertical mean value of the clusters was the same. The analyses reported in this section were conducted on trials in which the participants were performing the highest cluster task and responding to clusters with the same mean height.

Participants showed a statistically insignificant trend toward choosing the cluster on the right (53.8%, *CI* [49.8%, 57.9%], *p* = .066). Once again, all stimuli were counterbalanced across the left-right dimension; all further analyses are collapsed across whether the cluster was on the left or right side of the screen.

Sparsity

When a cluster with low sparsity and a cluster with high sparsity were paired, participants tended to indicate that the cluster with low sparsity was higher (79.2%, *CI* [70.8%, 86.0%], *p* < .001). This pattern was consistent across centering and dispersion manipulations. It held regardless of whether the clusters had low dispersion (low sparsity cluster chosen 78.3%, *CI* [65.8%, 87.9%], *p* < .001) or high dispersion (low sparsity cluster chosen 80.0%, *CI* [67.7%, 89.2%], *p* < .001) and whether the reference point was mean centered (low sparsity cluster chosen 81.7%, *CI* [69.6%, 90.5%], *p* < .001) or standard deviation centered (low sparsity cluster chosen 76.7%, *CI* [64.0%, 86.6%], *p* < .001).

Dispersion

When a low dispersion and high dispersion cluster were paired, participants tended to choose the cluster with high dispersion as having a higher vertical mean (70.8%, *CI* [61.8%, 78.8%], *p* < .001). Contrary to the reference point membership task, this pattern was consistent across centering method, along with sparsity manipulation. It held regardless of whether the clusters had low sparsity (high dispersion cluster chosen 66.7%, *CI* [53.3%, 78.3%], *p* = .013) or high sparsity (high dispersion cluster chosen 75.0%, *CI* [62.1%, 85.3%], *p* < .001) and whether the reference point was mean centered (high dispersion cluster chosen 66.7%, *CI* [53.3%, 78.3%], *p* = .013) or standard deviation centered (high dispersion cluster chosen 75.0%, *CI* [62.1%, 85.3%], *p* < .001).

Cluster Types

We examined which cluster type was preferred when the different types of clusters were pitted against one another. Once again, recall that clusters of type *A* had low dispersion and low sparsity, *B* had low dispersion and high sparsity, *C* had high dispersion and low sparsity, and *D* had high dispersion and high sparsity. Table Y below shows the results, collapsed across centering method (since the same pattern was found regardless of centering method). The preferred cluster type for the higher mean task ranks as follows: 1. *C*, 2. *D*, 3. *A*, and 4. *B*. Dispersion was the driving factor, with high dispersion clusters preferred. Sparsity was of secondary importance, with low sparsity clusters preferred over high sparsity clusters.

Clusters	Chosen Most	Percent	95% CI	p-value
A vs. B	A	78.3%	[65.8%, 87.9%]	< .001
A vs. C	C	67.7%	[53.3%, 78.3%]	.013
A vs. D	D	72.9%	[58.6%, 82.5%]	.001
B vs. C	C	78.3%	[65.8%, 87.9%]	< .001
B vs. D	D	75.0%	[62.1%, 85.3%]	< .001
C vs. D	C	80.0%	[67.7%, 89.2%]	< .001

Table Y. *Cluster types A (low dispersion, low sparsity), B (low dispersion, high sparsity), C (high dispersion, low sparsity), and D (high dispersion, high sparsity), compared in the highest cluster task. Percentage for most common choice, 95% confidence intervals, and p-value reported.*

Highest Point

The dispersion and sparsity manipulation also influence which cluster tends to have the highest overall point. Participants might simply be choosing the cluster with the highest overall point when deciding which cluster has the highest mean. To examine this possibility, we once again used GIMP software (GIMP Development Team, 2007) to hand code the pixel coordinates of the highest point in each cluster. We then determined which cluster of each pair had the highest point and whether on average participants were more likely to choose the cluster with the highest point. Participants chose the cluster with the highest point 85.0% of the time (*CI* [70.2%, 94.3%], *p* < .001). When comparing cluster types, the highest point pattern perfectly aligned with the cluster preferences (e.g., *A* tended to have a higher point than *B*, and *A* was preferred over *B*). Once again, because the overall highest point metric and the distribution of the cluster based on the dispersion and sparsity manipulation are strongly correlated, this design does not allow us to tease apart the individual contributions of each.[9]

Discussion

Dispersion was the primary driver in the highest cluster task, with clusters with high dispersion being seen as higher than those with low dispersion. Of secondary importance was sparsity: clusters with low sparsity were seen as higher than clusters with high sparsity. These manipulations aligned with which cluster tended to have the overall highest point. Reference point centering method wasn't influential in this task, which isn't surprising considering the reference is not relevant to the task.

Eye Movement Results

Fixations were calculated in the same way in Experiment 1B as in Experiment 1A.
We first examined overall differences in number of fixations and fixation duration between the two tasks (reference point membership and cluster height). A mixed effects model with a fixed effect for task and random intercepts for participant and stimulus (using Satterthwaite approximation for degrees of freedom) revealed that overall, participants had slightly more fixations on average in the cluster height task (*mean* = 4.75 fixations, *stdev* = 3.64) relative to the reference point task (*mean* = 4.59 fixations, *stdev* = 3.44; *t*(1900) = 2.28, *p* = .023). A similar mixed effects model with fixation duration as the fixed effect revealed that fixation durations in the reference point task (*mean* = 394.78 ms, *stdev* = 330.01) tended to be longer than those in the outlier task (*mean* = 347.04 ms; *stdev* = 257.36; *t*(8695) = 9.19, *p* < .001).

We also examined proportion of fixations to each of three ROI categories (cluster, reference point, and other). See Figure LL. A mixed effects model predicting proportion of fixations from the fixed effects of task (highest cluster vs. reference point membership) and type of ROI along with random intercepts for subject and stimuli (using Satterthwaite approximation for degrees of freedom) revealed significant simple effects of task for each the ROIs. Relative to the highest cluster task, participants in the reference point membership task tended to have a higher proportion of fixations to both the reference point ROI (*t*(5910) = 6.53, *p* < .001) and the "other" ROIs (*t*(5910) = 10.18, *p* < .001); in contrast, those in the highest cluster task tended to have a higher proportion of fixations to the cluster ROIs than those in the reference point membership task (*t*(5910) = 16.71, *p* < .001).

[9] We also examined the effect of highest point in each cluster when the means of the two clusters were *not* the same. When participants gave an incorrect response (e.g., indicated the left cluster had a higher mean when the right cluster actually had a higher mean), was it because the incorrect cluster had a higher point? Across 600 trials where the clusters were at different mean heights, participants made a mistake on 74 trials (12%). On those 74 trials, only 10 trials (14%) were on stimuli where the cluster with the *lower* mean had the overall *highest* point. Participants do not appear to primarily make errors due to reliance on highest point when there is a true difference in mean cluster height. We suspect that the driving effect of the overall 12% error rate on images with different means is due to user error (e.g., hit the wrong button or not appropriately following instructions).

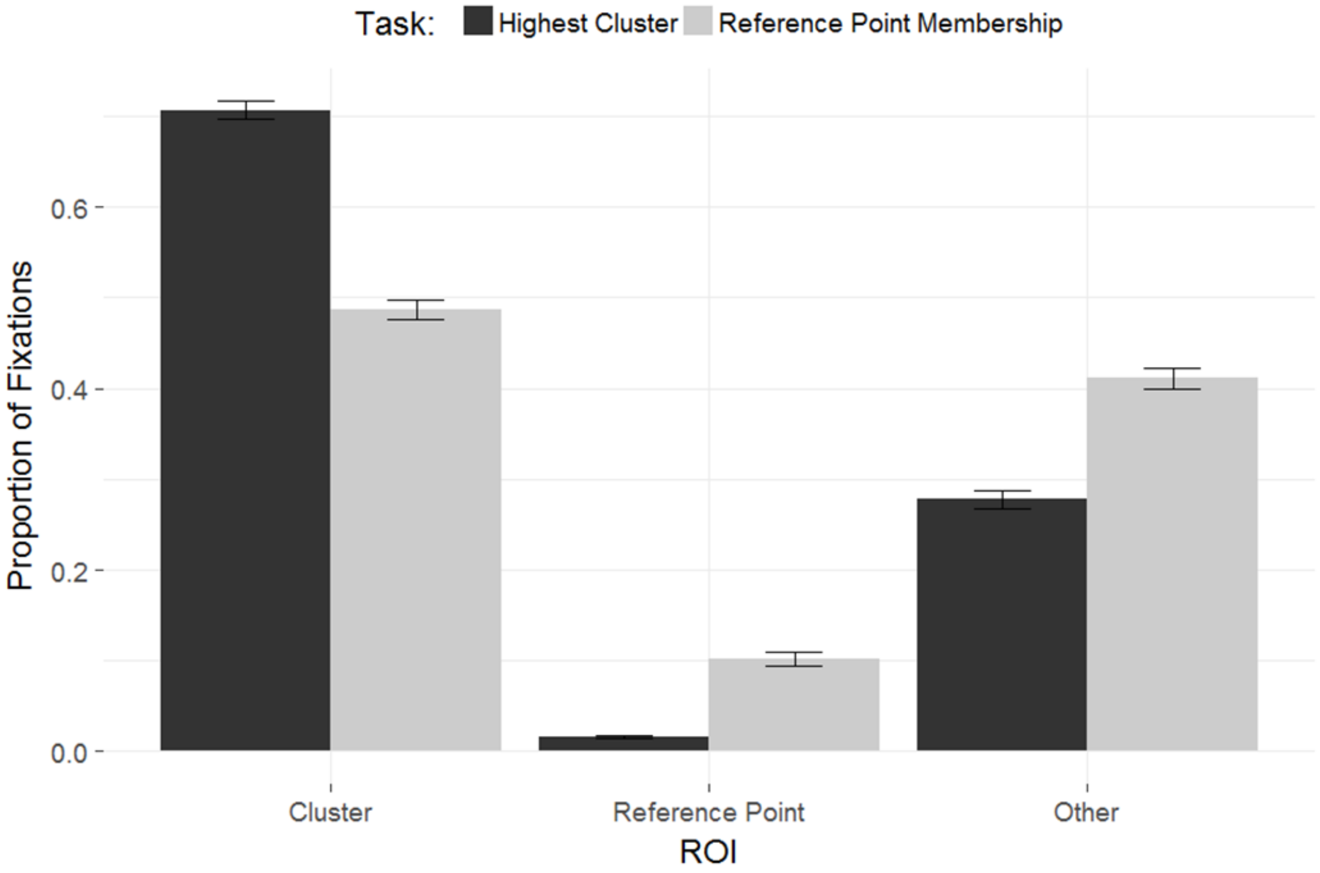


Figure LL. Proportion of fixations each subject made to regions of interest (ROIs) based on task (highest cluster or reference point membership) in Experiment 1B. Error bars represent standard error of the mean.

Experiment 2

Methods

Participants

Fifteen participants were recruited from students, faculty, and staff at the University of Illinois (5 males, *mean age* = 31.87 years, *stdev* = 10.84 years) and compensated $20 for their time. All participants were required to have at least one publication in a scientific journal and were tested for color vision deficiencies (24 plate Ishihara Test; Ishihara, 1972) and near vision acuity. The data from one participant was dropped and replaced due to colorblindness. Seven out of the fifteen participants also completed Experiments 1A and 1B.

Design

Type of plot (bar with standard error, bar with confidence intervals, dot, violin, box, or density), number of data points per variable (low or high), variance in the data (low or high), and p-value for the difference between the two variables (5 ranges) were manipulated within subjects. Participants interpreted the difference between two variables and gave confidence ratings.

Materials

Stimuli

All stimuli in the main task were created in R Software (R Development Core Team, 2008) from simulated data using the ggplot2 software package (Wickham, 2009). Twenty unique data sets were created by crossing 2 levels of number of data points per variable (n=25 or n=100), 2 levels of variance in the data (low or high), and 5 levels of p-values for an unpaired, two-tailed t-test of the difference between the two variables ($p < .001$, $.01 \leq p \leq .03$, $.04 \leq p \leq .06$, $.07 \leq p \leq .15$, and $.40 \leq p \leq .60$). Each data set contained two independent variables (labeled "A" and "B") drawn from Gaussian distributions. Whether the mean of A or B was higher was randomly determined. Each of the 20 data sets was plotted 6 ways: a bar plot with standard error of the mean, a bar plot with 95% confidence intervals, a jittered dot plot, a violin plot, a box plot (with 1.5 * Inter-Quartile Range error bars), and an overlaid density plot, leading to 120 different plots. See Appendix C for example stimuli.

Pre-Task Survey

Participants answered the following survey questions prior to completing the main task. How many years of experience do you have in research requiring statistical inference? What is your main field of research? How would you rate your statistical knowledge (5-point Likert scale)? What is the commonly acceptable value for statistical significance in your field? What does it mean if a t-test has a p-value of 0.01? What type of visualization or graph do you most frequently use when presenting your work?

Post-Task Survey

After completing the main task, participants answered whether they were familiar with all of the chart types and measures of error presented. If not, they were asked to identify which ones were not familiar.

Procedure

Experiment 2 was completed in the same environment with the same equipment as in Experiments 1A and 1B. Participants were tested for color vision deficiencies and near visual acuity, along with reporting their age and gender, prior to completing the study. Then participants responded to the pre-task survey questions, worked through the main task of viewing data visualizations, and responded to the post-task survey question.

In the main task, participants randomly viewed each of the 120 stimuli. The task was self-paced unless the participant did not advance within 10 seconds (at which time the program automatically advanced). After viewing each stimulus, the participant responded to two questions using mouse clicks: (A) Is the difference between Groups A and B statistically significant ($p < .05$)? (B) How confident are you in your response to (A)? (Likert scale from *1* to *5*, with *1* being not at all confident, *3* being moderately confident, and *5* being very confident). Eye tracking data was collected while stimuli were on the screen but not during a participant's response.

Behavioral Results

Survey Results

Pre-Task Survey

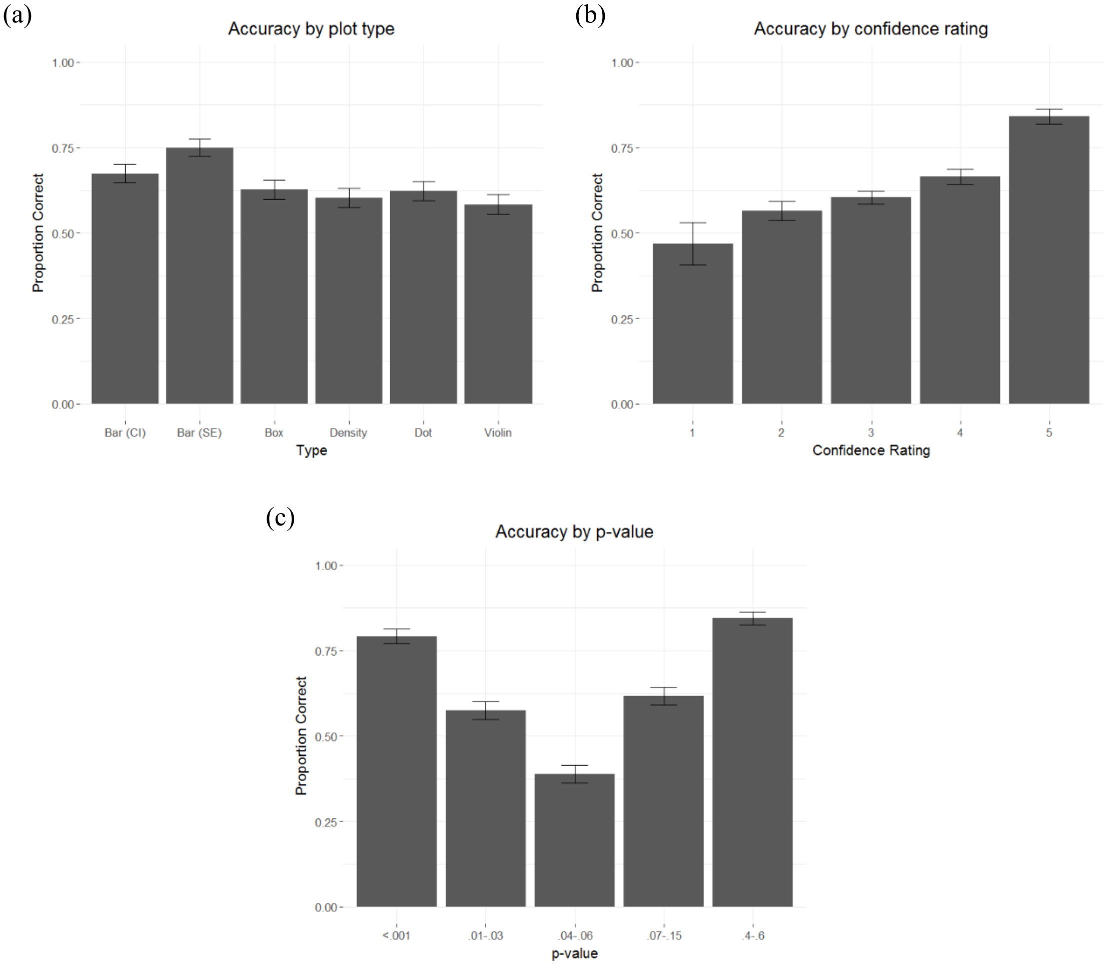
Participants had on average 8.7 years of experience in research requiring statistical inference (*stdev* = 5.9 years). Most participants primarily worked in a field related to psychology or neuroscience (13 participants); 1 each were in biology or engineering. The average rating on the statistical knowledge question was 3.6 (*stdev* = 0.8; 5-pt Likert scale with *1* for poor, *3* for moderate, and *5* for strong). Every participant sufficiently described the meaning of a p-value of 0.01 in a t-test. Bar graphs were the most commonly used visualizations (7 participants mentioned them). Two to three participants indicated they used scatterplots, line graphs, wave form plots, and/or violin plots. Pie charts, histograms, bivariate, and 3D plots were mentioned by a single participant. (Please note that a single participant was allowed to list more than one type of visualization s/he commonly used).

Post-Task Survey

Participants were least familiar with the box plot with inter-quartile range error bars (8 participants). Three participants each were unfamiliar with the violin or density plots. Three participants indicated they were familiar with all the visualization types used in the main task.

Visualization Task

While the eye movement behavior was of primary interest, we also looked at the behavioral responses to the task. We pulled plot type, p-value range, sample size, and variance together as fixed effects in a mixed effects model with a random effect for participant using the lme4 package in R software (Bates, Maechler, & Dai, 2011) to predict accuracy. See Figure NN for accuracy results. Accuracy was higher for bar plots than other plots overall ($Z = 3.26$, $p = .001$), with higher performance on bar plots with standard error of the mean than bar plots with 95% confidence intervals ($Z = 3.28$, $p = .001$). While we found a consistent numerical increase in accuracy, there were no significant differences in accuracy between adjacent steps of confidence levels (e.g., 1 relative to 2), with the exception of more accurate responses for those given a confidence rating of 5 relative to 4 ($Z = 3.48$, $p < .001$). Significant differences in accuracy were found moving between all adjacent ranges of p-values (all $Z > 4.9$, all $p < .001$). Accuracy was significantly higher for low sample sizes than high sample sizes ($Z = 5.93$, $p <.001$). No significant differences in accuracy performance were found for standard deviation differences.



Once again we labeled ROIs in each

Figure NN. Average accuracy performance by (a) plot type, (b) confidence rating, and (c) p-value range. Error bars represent standard error of the mean.

References

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision, 1*(4), 333-356.

Berg, D. & Itti, L. (2008). Memory, eye position, and computed saliency. *Journal of Vision, 8*(6), 1164.

Birmingham, E., Bichof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research, 49*(24), 2992-3000.

Card, S., Mackinlay, J. D., & Schneiderman, B. (Eds.). (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann.

Castelhano, M. S., Mack, M. L., & Henderson, J.M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology, 62*, 1-14.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*(5), 170-178.

Cleveland, W. S. (1993). *Visualizing Data*. Summit, NJ: Hobart Press.

Etemadpour, R., Olk, B., & Linsen, L. (2014). Eye-tracking investigation during visual analysis of projected multidimensional data with 2D scatterplots. *Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP)*, 233-246.

Findlay, J. M. & Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing.* Oxford, UK: Oxford University Press.

Foulsham, T. & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception, 36*, 1123-1138.

GIMP Development Team (2007). GNU Image Manipulation Program [Computer software manual]. GIMP 2.8.18. Retrieved from www.gimp.org.

Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics, 12*(19), 2316-2325.

Healey, C. G. & Enns, J. T. (1998). On the use of perceptual cues and data mining for effective visualization of scientific datasets. *Graphics Interface, 98*, 177-184.

Hegarty, M., Canham, M. S., & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *36*(1), 37-53.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*, 498-504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray and R. L. Hill (Eds.) *Eye Movements: A Window on Mind and Brain* (pp. 537-562). Oxford: Elsevier.

Hoffman, J. E. & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception and Psychophysics, 57*(6), 787-795.

Interrante, V. (2000). Harnessing natural textures for multivariate visualization. *IEEE Computer Graphics and Applications, 20*(6), 6-11.

Ishihara, S. (1972). Tests for colour-blindness: 24 plates edition. Tokyo: Kanehara Shuppan Co., Ltd.

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*, 194-203.

Jänicke, H. & Chen, M. (2010). A salience-based quality metric for visualization. *Computer Graphics Forum, 29*(3), 1183-1192.

Kim, Y. & Varshney, A. (2006). Saliency-guided enhancement for volume visualization. *IEEE Transactions on Visualization and Computer Graphics, 12*(5), 925-932.

Land, M. F., Mennie, N., & Rusted, J. (1999). Eye movements and the roles of vision in activities of daily living: Making a cup of tea. *Perception, 28*, 1311-1328.

Land, M. F. & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*, 3559-3565.

Lansdale, M., Underwood, G., & Davies, C. (2009). Something overlooked? How experts in change detection use visual saliency. *Applied Cognitive Psychology, 24*(2), 213-225.

Lleras, A., Rensink, R. A., & Enns, J. T. (2005). Rapid resumption of interrupted visual search: New insights on the interaction between vision and memory. *Psychological Science, 16*(9), 684-688.

Lleras, A., Rensink, R. A., & Enns, J. T. (2007). Consequences of display changes during interrupted visual search: Rapid resumption is target specific. *Perception and Psychophysics, 69*(6), 980-993.

Matzen, L. E., Haass, M. J., Divis, K. M., & Stites, M. C. (2016). *Using eye-tracking metrics and visual saliency maps to assess image utility.* Paper presented at the Human Vision and Electronic Imaging (HVEI) XXI.

Matzen, L. E., Haass, M. J., Divis, K. M., Wang, Z., & Wilson, A. T. (2017). Data visualization saliency model: A tool for evaluating abstract data visualizations. *IEEE Transactions on Visualization and Computer Graphics*.

Miccallef, L., Palmas, G., Oulasvirta, A., & Weinkauf, T. (2017). Towards perceptual optimization of the visual design of scatterplots. *IEEE Transactions on Visualization and Computer Graphics, 23*(6).

Michal, A. L. & Franconeri, S. L. (2017). Visual routines are associated with specific graph interpretations. *Cognitive Research: Principles and Implications, 2*(1), 20.

Michal, A. L., Uttal, D., Shah, P., & Franconeri, S. L. (2016). Visual routines for extracting magnitude relations. *Psychonomic Bulletin & Review, 23*(6), 1802-1809.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision, 11*(8).

Milosavljevic, M. N., V. Koch, C., & Rangel A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology, 22*(1), 67-74.

Munzner, T. (2014). *Visualization Analysis and Design*. Boca Raton, FL: CRC Press.

Nothelfer, C., Gleicher, M., & Franconeri, S. (2017). Redundant encoding strengthens segmentation and grouping in visual displays of data. *Journal of Experimental Psychology: Human Perception and Performance*.

R Development Core Team (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://r-project.org (ISBN 3-900051-07-0).

Rayner, K., Rotello, C. M., Stewarrt, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movments when looking at print advertisements. *Journal of Experimental Psychology: Applied, 7*(3), 219-226.

Veas, E. E., Mendeaz, E., Feiner, S. K., & Schmalstieg, D. (2011). Directing attention and influencing memory with visual saliency modulation. *Proceedings of the SIGCHI Conference on Human Factors in Computer Systems*, 1471-1480.

Wickham (2009). *ggplot2: Elegant graphics for data analysis*, Dordrecht New York: Springer. Retrieved from http://ggplot2.org (ISBN 978-0-387-98140-6).

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1*(2), 202-238.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception & Performance, 15*(3), 419-433.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

APPENDIX F: DATA VISUALIZATION SALIENCY MODEL: A TOOL FOR EVALUATING ABSTRACT DATA VISUALIZATIONS[10]

[10] Matzen, L. E., Haass, M. J., Divis, K. M., Wang, Z., & Wilson, A. T. (in press). Data Visualization Saliency Model: A Tool for Evaluating Abstract Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics*.

Laura E. Matzen, Michael J. Haass, Kristin M. Divis, Zhiyuan Wang, and Andrew T. Wilson

Abstract—Evaluating the effectiveness of data visualizations is a challenging undertaking and often relies on one-off studies that test a visualization in the context of one specific task. Researchers across the fields of data science, visualization, and human-computer interaction are calling for foundational tools and principles that could be applied to assessing the effectiveness of data visualizations in a more rapid and generalizable manner. One possibility for such a tool is a model of visual saliency for data visualizations. Visual saliency models are typically based on the properties of the human visual cortex and predict which areas of a scene have visual features (e.g. color, luminance, edges) that are likely to draw a viewer's attention. While these models can accurately predict where viewers will look in a natural scene, they typically do not perform well for abstract data visualizations. In this paper, we discuss the reasons for the poor performance of existing saliency models when applied to data visualizations. We introduce the Data Visualization Saliency (DVS) model, a saliency model tailored to address some of these weaknesses, and we test the performance of the DVS model and existing saliency models by comparing the saliency maps produced by the models to eye tracking data obtained from human viewers. Finally, we describe how modified saliency models could be used as general tools for assessing the effectiveness of visualizations, including the strengths and weaknesses of this approach.

INTRODUCTION

Vision is the dominant sense for humans [2], with researchers estimating that over 50% of the brain is involved in processing visual information [1,39]. Given how heavily most humans rely on vision to navigate and understand the physical world, it is no surprise that visualizations are a common tool for helping people to navigate through information. Visualizations leverage the capabilities of the human visual system and can provide users with a natural way to explore and comprehend large amounts of information. However, visualizations can also be confusing and misleading, particularly for complex, multidimensional data sets that do not have a natural visual representation.

Evaluating the effectiveness of visualizations can be very challenging [10,30]. Ideally, visualizations would be evaluated with well-designed user studies, but these are not always possible (e.g. if the designer does not have access to the end users) and can also be expensive and time consuming. It would be useful for designers to have more evaluation tools that can be deployed rapidly and iteratively during the design process to assess visualizations prior to conducting a user study. Prior work has suggested that visual saliency models could be one such tool [26,38].

Visual saliency models assess the visual features of an image to predict which areas of that image will draw a viewer's attention. Saliency models are typically inspired by the structure and function of the human visual cortex. The models take an input image and generate a saliency map that predicts which regions of the image will be most likely to draw a human viewer's attention [24]. There are a variety of metrics that can be used to assess the performance of the models by comparing the saliency maps to human fixation data recorded via eye tracking [4,7,8]. Saliency models have been the subject of a great deal of research in the fields of cognitive science and computer vision, and they could prove useful to visualization designers as well. Since data visualizations make use of the human visual system to convey information, evaluation techniques that are rooted in neural processes could provide useful, generalizable metrics.

It is important to note that saliency models' predictions of where viewers will look are based only on the physical properties of the visual stimulus. They are models of what is known as *bottom-up* visual attention. In real-world tasks, a viewer's eye movements are also guided by *top-down* visual attention, which is influenced by the viewer's goals, expectations, and experience [12,43,46]. In the brain, these two processes operate in parallel. Bottom-up visual attention is drawn to regions of a stimulus that are distinct from things around them in terms of their basic visual features (e.g. contrast, color, motion), and top-down visual attention is allocated voluntarily based on the viewer's task and prior knowledge. Regions with high bottom-up saliency may or may not be relevant to the viewer's task and goals, so there is a constant interplay between the two neural systems that guide visual attention and eye movements [41].

When a saliency model is applied to an image, it produces a map that predicts which regions of the image are most likely to draw the viewer's bottom-up attention. In the context of data visualizations, this could allow designers to assess whether or not their design will draw attention to the most important information [26]. In other words, saliency maps provide designers with a metric of how well bottom-up attention and top-down goals will overlap for the application that the designer has in mind. From the perspective of a person using a visualization, a strong overlap between visual saliency and important features will allow the user to complete tasks faster and more efficiently, minimizing distraction from unimportant information.

Although generating saliency maps for data visualizations could provide a useful and widely applicable evaluation metric, there is a substantial obstacle to this approach. The existing models of bottom-up visual saliency were designed for images of natural scenes, and the visual and spatial properties of natural scenes can be quite different from those of visualizations. While saliency models can generate reasonable predictions of where people will look in scene-like visualizations (i.e., visualizations that resemble photographs) [38], these models typically underperform for abstract visualizations [18].

This is a disadvantage for existing saliency models, but it raises the possibility that these models can be modified to better account for patterns of attention in data visualizations. The differing nature of visualizations and natural scenes also presents opportunities to incorporate some information about top-down attention into saliency models. In the context of natural scenes, top-down attention is highly task- and situation- dependent, making it very difficult to model in any generalized way. This is the reason that most existing saliency models take only bottom-up attention into account. However, in the context of data visualizations, the visual features and their placement within the scene are selected by a designer in support of a particular goal or goals. A designer is structuring the image in order to convey information, so the visual features that the designer selects encode top-down information in a way that the features of a natural scene do not. Visualizations are also typically “born digital,” unlike images of natural scenes, making it easier to isolate distinct elements (such as individual data regions or text regions) and infer their importance from a top-down perspective.

In this paper, we explore why existing saliency models underperform for abstract data visualizations. We identify the visual and structural features of visualizations that are incompatible with the existing, scene-based visual saliency models. We then discuss the development of a modified saliency model that addresses these features and incorporates new information based on top-down attention, allowing it to make more accurate predictions of which regions of a visualization will draw a viewer’s attention. We outline the features of the Data Visualization Saliency (DVS) model and compare its performance to a set of existing saliency models. Finally, we discuss how the DVS model could be used as an evaluation tool during the process of designing a visualization, allowing designers to rapidly assess how various design choices affect the saliency of different parts of a visualization.

1 Evaluation of Existing Saliency Models

There are numerous bottom-up saliency models that have been developed to predict where people will look in natural scenes. Many of these models are based on the neurophysiology of human and other primates’ visual systems [3]. They select visual features that are known to elicit neural responses in the visual cortex, such as luminance, hue, contrast and orientation. The feature maps are often created at multiple scales of image resolution, filtered, and then combined to produce a master saliency map. The performance of saliency models is assessed by comparing the saliency maps produced for a range of stimuli to eye tracking data obtained from human viewers looking at the same stimuli.

The MIT Saliency Benchmark project [7] keeps a running scoreboard for author-submitted models, showing how well they predict human fixations on benchmark image sets. The project includes two sets of benchmark images and corresponding fixation data recorded from human viewers. The project has also established eight metrics for assessing the match between saliency and fixation maps [8]. A full discussion of each metric is outside of the scope of this paper (see [8,18] for more detailed descriptions), but each metric is briefly described below.

Three of the eight metrics are location-based, meaning that they assess how well saliency maps predict the location of human fixations in an image. All three of the location metrics are based on the concept from signal detection theory of the Area under the Receiver Operating Characteristic (ROC) Curve, or AUC. The three variants of this approach are AUC-Judd, AUC-Borji, and shuffled AUC (sAUC). Scores range from 0 to 1 with 1 being the optimal score and 0.5 representing chance performance. The key differences between these three metrics lie in how they calculate true and false positives. For example, AUC-Borji uses a uniform random sample, while the sAUC, which was developed specifically for assessing saliency models, samples in a way that penalizes models that are biased toward the center of the image [8].

Four metrics are based on comparisons of the distribution of fixations across an image to the distribution of saliency in a saliency map. These metrics are called the similarity metric (SIM), Earth Mover’s Distance (EMD), Pearson’s Correlation Coefficient (CC), and Kullback-Leibler divergence (KL). The SIM metric treats the fixation and saliency maps as histograms and assesses their overlap. Scores range from 0 to 1, with 1 indicating perfect overlap. False negatives are highly penalized under the SIM metric. The EMD computes the cost of transforming one map to the other. If two distributions are identical, the EMD is zero, so lower scores represent better performance. CC measures how correlated the two maps are, penalizing false negatives and false positives equally. A score of 1 represents a near-perfect correlation between the saliency and fixation maps. KL is an information theoretic measure that assesses the information lost when the saliency map is used to approximate the fixation map. A score of zero is optimal, so lower scores represent better performance for the saliency map. The KL metric is particularly sensitive to zero values, so sparse saliency maps are penalized with high KL scores [8,18].

Finally, the Normalized Scanpath Saliency (NSS) is a value-based metric. It standardizes the saliency map and then computes the average saliency at locations that were fixated. When the NSS score is greater than 1, that indicates that the fixated locations had significantly higher saliency than other locations in the image [8,18].

The visual saliency modelling community has not settled on any single metric for evaluating model performance. We feel it is important to consider at least one metric from each category (value, location, distribution) because corner cases may be easier to identify when comparing results from metrics in different categories. For consistency with prior publications, and in hope of compatibility with future investigations, we provide results for all of the eight metrics in the evaluations discussed below.

Saliency models are generally trained and tested using images of natural scenes. One of the two sets of benchmark images provided by the MIT Saliency Benchmark, the MIT300 set, consists of 300 images of indoor and outdoor scenes. The other dataset, CAT2000, consists of 2000 training and 2000 test images organized into 20 categories. Of the 20 categories, 15 are comprised of images of natural scenes. These are either photographs or manipulations of photographs, such as inverted or low resolution images. The remaining five categories contain images that are more abstract, such as cartoons, sketches, and fractals.

In a prior study [18], we sought to assess the performance of existing visual saliency models on data visualizations, a category that is not represented in the CAT2000 benchmark. We selected three saliency models that spanned a range of performance on the CAT2000 benchmark: the Itti, Koch and Niebur model [25], the Boolean Map Based Saliency model (BMS) [48], and the Ensembles of Deep Networks Model (eDN) [45]. We measured the performance of each of the selected models on a set of 184 data visualizations drawn from the Massachusetts (Massive) Visualization Data Set (MASSVIS) [6]. These were common types of data visualizations (bar charts, pie charts, etc.) that had corresponding eye movement data from human viewers. For each model, saliency maps were generated for each visualization and compared to the fixation maps using the eight metrics discussed earlier.

This analysis found that all three saliency models generally performed worse on the visualizations than on the images from the CAT2000 data set. The BMS model, which is one of the highest performers on the CAT2000 benchmark, performed significantly worse on data visualizations relative to the CAT2000 images for 6 of the 8 evaluation metrics. The eDN model had significantly worse performance according to five of the eight metrics. Interestingly, the Itti model, which has the lowest average performance of these three models on the CAT2000 set, performed best on the data visualizations. However, it still performed significantly worse on data visualizations than on the CAT2000 images according to four of the eight metrics.

A simple example of the models' underperformance on visualizations is shown in Figure 1, which provides one example from the MASSVIS set with corresponding fixation and saliency maps. Note that most of the fixations (Panel B) were devoted to the text labels for the bar graph. In contrast, the three saliency models tend to predict that viewers will fixate on the bars themselves due to their high contrast, sharp edges, and central location in the image. The reasons for this mismatch are outlined in more detail below.

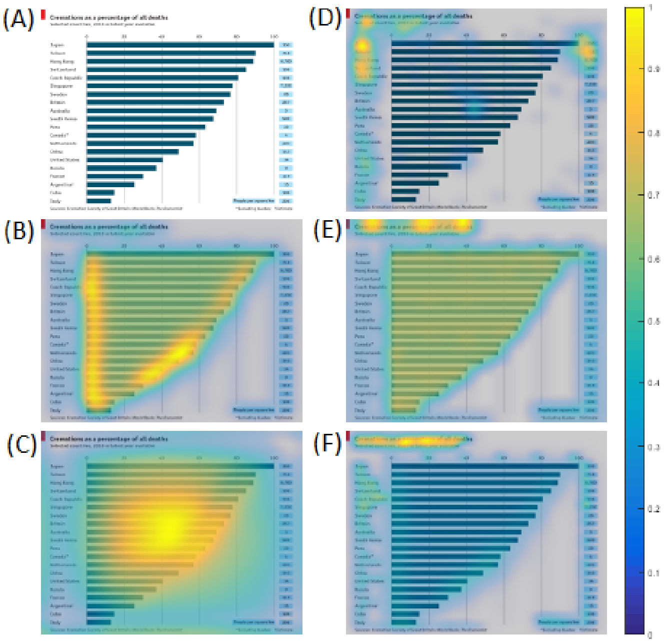


Figure 1. Fixation map and saliency maps generated by different models for an image from the MASSVIS set. (A) the original data visualization; (B) fixation map from Borkin et al. [5]; (C) Itti model; (D) BMS model; (E) eDN model; and (F) DVS model.

2 Differences in Visual Properties of Data Visualizations and Natural Scenes

It is clear from the analysis outlined above that existing visual saliency models are inadequate for predicting where people will look in abstract data visualizations. Models that generally perform quite well on natural scenes, and even somewhat abstract imagery such as cartoons, performed significantly worse on common types of data visualizations. We hypothesize that the reason for this poor performance is that the spatial scales and visual features used by the saliency models are inadequate for data visualizations.

2.1 Spatial Scales

Each of the models discussed above (Itti, eDN and BMS) follows a common approach. First, for each type of visual feature used by the model, "interestingness" maps (or "conspicuity maps," after Itti et al. [25]) are computed at one or more resolutions. Second, the individual feature maps are combined into an overall attention map and then into a saliency map.

As an example, the Itti model operates on multiple spatial scales by constructing a Gaussian pyramid from the input image. At each level of the pyramid, a Gaussian smoothing function is applied and the image is subsampled by a factor of two, creating a smaller, smoothed version of the image, as shown in Figure 2. A feature map is computed for each level, and then the feature maps are compared across levels of the Gaussian pyramid. Image regions with the greatest difference in feature values across scales are assigned higher saliency values than regions with smaller differences across scales. This comparison process is the model's implementation of the center-surround neural activation properties of the human visual system.

Although this approach works relatively well for natural scenes, the spatial properties of data visualizations are quite different. Many of the elements in data visualizations (glyphs, lines, text) are quite small, and visualizations are likely to have a higher proportion of small but important variations than natural scenes. The smoothing and subsampling process results in the loss of these small details. For example, text becomes blurry at the first level of smoothing, leading to minimal differences between the levels of the Gaussian pyramid when the visual features of the text are compared across scales. This results in low saliency values for text even though text typically receives a high proportion of fixations [37].

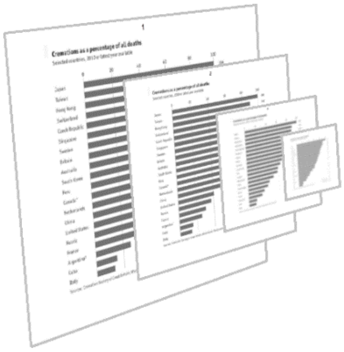


Figure 2. Example of a Gaussian pyramid with four levels of smoothing and resizing.

Another problematic aspect of the existing saliency models is that many of them resize the input image to a standard size as their first step. For example, the BMS model begins by resizing the input to be exactly 600 pixels wide. Similarly, the reference implementation of the eDN model resizes its input to a resolution of 512x384. While this makes the computation go quickly, it also tends to blur text into unrecognizability and obliterate fine contours completely. This is a particular problem for visualizations since the meaningful elements of many data representations (line charts, box charts, some geographic maps and weather diagrams) are nothing but fine contours.

2.2 Visual Features

While the way in which the models combine their feature maps is fundamentally similar, they differ in terms of the specific visual features used to create the feature maps. The Itti model computes center-surround operations on intensity, orientation and color channels and combines them to create the attention map. It computes four color maps (red, green, blue and yellow) using RGB pixel values. The eDN model uses a support vector machine trained over many randomly constructed hierarchical features [42]. These features operate variously on RGB, YUV and grayscale images. The BMS model uses exactly one feature – connected regions. It computes these regions at multiple intensity thresholds using the channels of the CIE LAB color space.

2.2.1 Color

Since all three models compute some or all of their features over color channels, we believe that the color space chosen for these computations is particularly important. In our assessment of the three saliency models using the MASSVIS images, we noted that the models often assigned low saliency values to bright red regions, causing discrepancies between the saliency maps and the map of human fixations. We believe that this mismatch is driven by the fact that human color perception is very different from the way colors are created on paper or on an electronic display. This difference manifests in two ways. First, color spaces such as RGB or CMYK that are defined by the properties of an output device are perceptually non-uniform. That is, adding 0.1 to the red component of a color produces a larger perceived difference for some colors than for others. Second, the different "channels" of human color perception are not independent as they are in the case of display primaries. That is, adding redness while keeping luminance constant may change perceived luminance.

The YUV color space uses a luminance + chrominance representation of color that it is designed to permit efficient compression while minimizing artifacts. From the perspective of perceptual uniformity, YUV is an improvement over RGB but still leaves much to be desired. In order to do color arithmetic in a way that yields perceptually comparable results, it is advisable to work in a color space like CIE XYZ or CIE LAB [14]. The XYZ model operates with the tristimulus values obtained from the color-sensitive cones in the retina. The LAB model transforms these into a luminance channel (L) and two color-opponent channels (A and B) that agree with current thinking about the way color is processed in the brain. The LAB model has the additional advantage of being perceptually uniform. Adding 0.1 to a color component produces a change that appears to the observer to be of the same magnitude regardless of where it is in the color space. As a result, feature maps computed over different channels in the color space have values that can be meaningfully compared with one another.

2.2.2 White Space

A crucial difference between visualizations and natural scenes is the presence of white space. The real world is cluttered and natural scenes tend to have information (in the Shannon sense) absolutely everywhere. Synthetic scenes do not: they often contain large areas of uniform, untextured color. Some of these may be objects, but some are simply blank areas. Distinguishing between the two is a challenge. In either case, feature-based saliency models may have trouble "seeing" these regions since they will only be detectible a very coarse scale.

The spatial distribution of figures relative to the background is also quite different for abstract data representations than for physical objects. Many saliency models use a center weighting. This works well for photographs, where objects of interest are often centered. However, it may not be appropriate for visualizations, where meaningful information can appear in any spatial location and is often deliberately distributed across the entire image.

2.2.3 Text

As mentioned above, text in data visualizations receives a great deal of attention from viewers. In prior work, we have found that people viewing data visualizations while performing memory or free viewing tasks devote a disproportionate amount of attention to regions containing text. For example, in one dataset, an average of 60% of the participants' fixations fell in regions containing text, relative to 30% in regions containing visual representations of data [37]. In general, participants were highly likely to view regions containing text and to view them relatively early in the trial.

There are several causes for the high proportion of fixations devoted to text in visualizations. In general, literate people's attention is automatically drawn to text [28,33,35]. In data visualizations, text often provides context and details that are necessary for understanding the data. For example, our prior work found that participants are likely to refer to text-containing regions such as the legend and data labels multiple times as they view the visualization [37]. Finally, reading text requires numerous fixations. Under normal conditions, the estimated visual span for reading is about 10 letters [31]; words presented in peripheral vision cannot be resolved due to low visual acuity and crowding.

While text draws attention and necessitates many fixations, it is not included as a feature in most saliency models. The models are tailored to and/or trained on images of natural scenes, which rarely contain text. Our analysis of the performance of existing saliency models on data visualizations indicates that assigning appropriate levels of saliency to text is one of the key areas in which their performance could be improved.

3 The Data Visualization Saliency Model

Existing saliency models fall short for data visualizations, but our analysis of several models revealed concrete steps that can be taken to adapt them to this domain. We have developed the Data Visualization Saliency (DVS) model[‡‡‡], which builds on the strengths of existing models while extending their capabilities to account for the visual features and spatial scales that are common in data visualizations. The two primary components of the current implementation of the model are a modified version of the Itti model and a text recognizer, which allows us to detect one of the key features of visualizations that is missed by current models. The DVS model combines the outputs of the modified Itti model and a text map to produce saliency maps that are specialized for data visualizations.

3.1 Modified Itti Model

We took as a starting point the Itti, Koch and Niebur saliency model [25] as implemented in the Graph Based Visual Saliency (GBVS) toolbox [20,21]. Of the existing models that were tested with data visualizations, this model had the highest performance [18]. The authors of the GBVS saliency model note that the original Itti model uses a simple color opponency representation based on RGB values. As discussed above, using the RGB color space is suboptimal, particularly in the case of data visualizations, where colors are chosen deliberately by a designer. To better approximate human visual perception, we modified the original algorithm by transforming the representation of the input images into CIE LAB color space. This change is likely to improve the model's performance for all types of imagery, but it is particularly important for visualizations, in which colors are deliberately selected to convey information.

3.2 Text Saliency Map

As discussed above, viewers devote a great deal of attention to text in data visualizations, yet text is not highlighted in existing saliency models. Although text regions often have high contrast, they tend to be small. The high-frequency details of text are lost when an input image is resized or smoothed. This leads to few differences across the levels of the Gaussian pyramid, and the text regions are not identified as being salient. To account for viewers' tendency to fixate on text in visualizations, we developed a text saliency model that could be combined with the modified Itti model. Attention to text is primarily driven by top-down visual attention, since people expect text to contain meaningful information. By incorporating this feature into our model, we are taking a step towards a saliency model that takes both bottom-up and top-down attention into account.

Our goal was to build an algorithm that computes the likelihood of belonging to a text region for each pixel of an input visualization image. Text detection is a popular challenge in the computer vision literature, and numerous successful models and algorithms have been developed in this domain. Detecting text in visualizations is a relatively easy task compared to detecting text from photos of real-world scenes. The method we detail below is essentially a combination of various classic text detection techniques. However, instead of producing a binary output, like traditional text detection algorithms, this method produces a continuous, probabilistic output that can be incorporated into a saliency map.

We used a common approach in the text detection literature, which is to extract Maximally Stable Extremal Regions (MSER) [36] as candidate text regions, and then to apply various text-diagnostic features to filter out the non-text candidates (e.g., [11,17,40]). The MSER algorithm detects connected, homogeneous ("maximally stable") regions of pixels. Because text almost

[‡‡‡] Available at: https://github.com/mjhaass/DataVisSaliency.git

always has uniform color and each letter in English is connected (in the sense that each "stroke" is connected to all other strokes in the same letter), English letters should be detected as MSER regions (i.e., the miss rate should be very low).

In order to exclude MSER regions that are not text, all detected MSER regions went through a filtering process based on simple properties of these regions, such as aspect ratio [11], Euler number [17,40], and solidity [17]. As an example, for most fonts of English letters and Arabic numerals, the height-to-width ratio should be less than 4 and greater than 1/3, so the aspect ratio of the bounding box of MSER regions was restricted to this range [11]. Finally, the data was filtered based on stroke width variation [17,32]. The variability of each MSER component's stroke width was compared to its mean stroke width. If the relative variability was too large, the region was filtered out (since letters and digits have relatively small stroke width variations).

After the above filtering, the remaining MSER regions had a relatively high likelihood of being letters or digits. In order to quantify this likelihood, we computed three text-diagnostic edge features on these regions (using simplified versions of the algorithms proposed by [34]). We took the bounding box of each MSER region and computed these features on the image patch defined by the bounding box. The three feature values were then summed together to form the raw "text saliency" score.

The first feature was based on the magnitude of the image gradient. For each image patch (i.e., each MSER region), the image gradient was computed on the grayscale transformation of the original colored patch. The mean gradient magnitude *μ(G)* and the standard deviation *σ(G)* of gradient magnitude were computed with *P* as a scaling constant:

$$F_1 = P\frac{\mu(G)}{\sigma(G)} \tag{1}$$

This feature is akin to a signal-to-noise ratio. In most scenarios, text strokes appear on a highly uniform background; the variability of the gradient magnitude is low but the text edges lead to high gradient magnitudes. This ratio should be high when the image patch contains text.

The remaining features were based on the edges in an image patch. For each MSER region, the Canny edge detection algorithm [9] was used to compute an "edge image" for each color channel of the image patch as represented in the CIE LAB color space.

The second feature attempts to capture a specific topological characteristic of text. Most text characters have either multiple strokes that intersect each other or curved strokes so that a vertical or horizontal "scan line" may cross the character body more than once. Since each stroke produces two edges, such "scan lines" will very likely cross the edges of the character more than twice. Therefore, the frequency of multiple-crossing by a scan line that scans horizontally and vertically is diagnostic of text. The higher the frequency, the more likely the image patch contains text. Formally, this feature can be given as

$$F_2 = Q^{(\sum_{i=1}^{H} f(cn_i) + \sum_{j=1}^{W} f(cn_j)) \big/ (W+H)} \tag{2}$$

where *W* and *H* are the width and height of the image patch in pixels, cn_i and cn_j denote the number of crossings for a specific scan line (vertical and horizontal respectively) and the edges in the image patch, and *f(x)* is a function that returns 1 when *x* is larger than 2 and 0 when *x* is equal to or less than 2. The constant *Q* is for scaling and weighting purposes. Using an exponential function with base *Q* increases the feature's sensitivity to higher multiple-crossing event counts and reduces sensitivity to small counts (which can occur randomly in non-text regions).

The third feature was based on a more straightforward characteristic. Text strokes usually produce two parallel edges, so that the number of crossings between a vertical or horizontal scan line and the text edges is often an even number. Hence the third feature can be defined similarly to the second one:

$$F_3 = R^{\left(\sum_{i=1}^{H} g(cn_i) + \sum_{j=1}^{W} g(cn_j)\right) \big/ (W+H)} \tag{3}$$

where *g(x)* returns 1 if *x* is an even number and 0 if it's odd. In the current implementation, the values of the scaling constants are *P* = 2.5, *Q* = 4, *R* = 1.22.

The text-specific feature values were normalized, combined, and treated as an index of probability of text in each region. The combined value of the three features was assigned to the pixel at the center of the region. This procedure was computed at different scales on the original image in order to enhance the method's sensitivity to smaller and larger fonts. The text saliency indices computed at each scale were re-scaled to the original image size and then combined by averaging. This raw text saliency map was then processed with Gaussian smoothing to simulate the randomness in the exact locations of human fixations.

3.3 Linear Combinations of the Model Components

Because there is insufficient data to inform how to best combine the text saliency map and the modified Itti saliency map, we opted for the simplest approach: a linear combination. Formally, the DVS model's saliency map S for a given visualization is computed as follows:

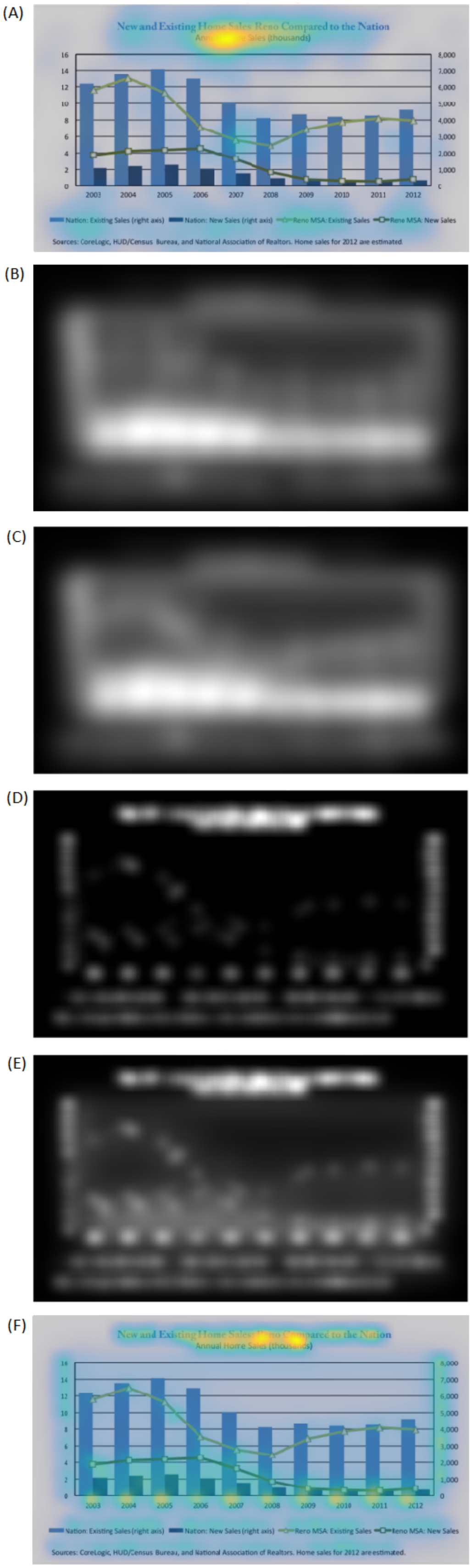
$$S = \frac{(I + w * T)}{(1 + w)} \qquad (4)$$

where *I* is the saliency map given by the modified Itti saliency model, and *T* is the text saliency map. The parameter *w* determines the relative weight between *I* and *T*. Both *I* and *T* are linearly scaled to have values ranging from 0 to 1 before combination. The denominator, (1 + *w*), produces a weighted average to maintain the overall saliency scaling from 0 to 1. Thus, for each data visualization image, a series of saliency maps based on a series of weight values can be generated. In order to choose an appropriate weight for the text saliency map, we systematically manipulated linear combinations of *I* and *T* and compared the resulting saliency maps to eye tracking data from the MASSVIS project [5]. The MASSVIS data set provides 393 data visualization images and corresponding fixation data. Thirty-three participants viewed the images while trying to memorize them for a later test. One visualization was excluded from our evaluation because it had an irregular size (less than 128 pixels wide) that is incompatible with the Itti saliency model. Thus, saliency maps and performance metrics were computed on the remaining 392 images.

We were primarily interested in how the average value for each of the eight MIT Saliency Benchmark evaluation metrics changed as a function of relative weight w between the modified Itti saliency map *I* and the text saliency map *T*. When *w* = 0, the saliency map *S* is just the modified Itti map; similarly, when *w* → ∞, *S* is equivalent to the text saliency map *T*. If the bottom-up saliency component captured by the modified Itti map *I* and the text-directed attention captured by *T* do complement one another, at some nonzero value of *w*, the combined map *S* should provide higher performance than either *I* or *T*. In other words, the performance-relative weight function should have a maximum point. Because of the differences in the nature of these metrics, we expect these functions to have different maximum points. Our goal was to find a reasonably good estimate of the window of *w* values in which the function reaches maximum for each of the eight metrics. Figure S1 in the Supplemental Materials plots each metric as a function of the weight parameter.

Notably, the baseline performance for the text saliency model was better than the baseline performance of the modified Itti model for six of the eight metrics (the SIM and KL metrics were the exceptions, likely because the text saliency maps include large regions that contain only zeros, and both of these metrics heavily penalize false negatives). The preference for the text saliency model is consistent with prior analyses showing that viewers disproportionately devote their attention to the text in the MASSVIS images [37]. Modelling only the text regions is a reasonable approximation for where people look in this particular data set and task. However, across all eight metrics, the linear combination of the modified Itti model and the text saliency model produced significantly higher matches to the human fixation data than either model alone.

The weight functions for each metric exhibit different shapes, reaching their maxima at different weight values. This aspect of the data was expected and supports the assertion that the eight metrics emphasize different aspects of the performance of a saliency model. There is no objectively optimal choice of the text saliency map weight, since no unique weight value optimizes all metrics of performance. In our experience, the choice of weighting factor typically causes performance results to fall into one of three categories; *under fit*, where performance increases proportionally to the weighting factor,



acceptable, where the performance is stable, or changes very slowly with changing weighting factor, and *over fit*, where performance may increase, but the gain on a given test case is likely not to transfer to another test case. Figure S1 shows that at least four of the performance metrics are approaching an asymptotic limit as the weight factor value approaches 2. To reduce the risk of over fitting, we chose to use a weight of 2 in the following analyses. Users of the DVS model can easily adjust this weight, if desired.

Figure 3 shows a representative example of the differences between the DVS model and the original Itti model. Additional examples are provided in the Supplemental Materials. The top panel of Figure 3 shows a data visualization from the MASSVIS set with overlaid fixation data (A). The remaining panels show the saliency maps produced by the original Itti map (B), the modified Itti map (C), the text saliency map (D), and the final, weighted DVS saliency map (E). Finally, the bottom panel (F) shows the DVS map overlaid on the original image, using the same color scale as the fixation map, allowing for a visual comparison of the two. Note that the original Itti map identifies the lower portion of the bar chart as the most salient region. The differences between the original Itti map and the map with the modified color space are subtle, but the modified model appears to do a better job of picking out the line graphs. The text saliency map correctly identifies all of the text regions in the image, but also has a few false alarms to features in the data, such as the data points on the line graphs. The DVS saliency map indicates that the title is highly salient, as is the lower part of the chart and the labels at the bottom of the chart. This corresponds well to the actual distribution of viewers' fixations.

3.4 Comparing the DVS Model to Existing Saliency Models

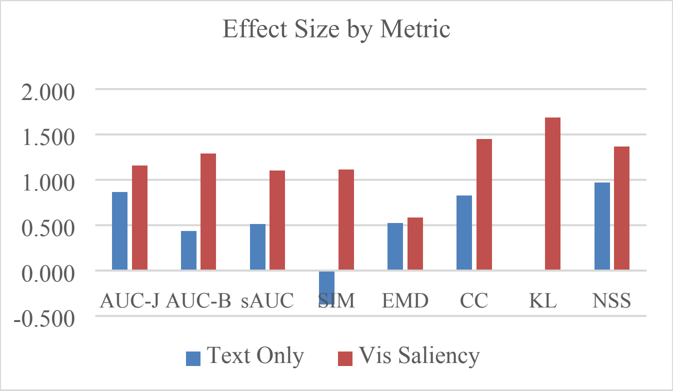
Once the weights in the DVS model had been optimized, the performance of the final model was compared to the original Itti model (as implemented in the GBVS toolbox), the BMS model, the eDN model, and to the text saliency maps alone. All of the models were used to generate saliency maps for 392 data visualizations from the MASSVIS dataset that had corresponding eye tracking data (as before, one visualization was excluded because its dimensions were incompatible with the Itti model). The saliency maps were compared to the eye tracking data using the eight metrics that are used by the MIT Saliency Benchmark. A one-way ANOVA was run for each metric, showing that there was a significant difference in the performance of the five models on all eight metrics (all *F*s > 44.69, all *p*s < 0.001).

Table 1 shows the percentage of improvement for the final, weighted DVS model relative to the Itti, BMS, eDN, and text saliency models on all eight metrics. The DVS model offered a substantial improvement in performance over the other models. Since the DVS model is based on the Itti model, we paid particular attention to how the components of the DVS model performed relative to the original Itti model. Figure 4 shows the effect size, using Glass's delta, for the improvement in performance for the text saliency maps and the final DVS model relative to the original Itti model. Notably, for all of the metrics other than EMD, the improvement in performance over the original Itti model was larger than one standard deviation. Performance also improved for the EMD metric, but the magnitude of the improvement was smaller. Finally, we used paired t-tests to assess whether or not the DVS model, as implemented with a weighting of 2, performed better than the text saliency maps alone. The KL metric was excluded from this analysis because its high sensitivity to zero values produced abnormally large scores for the text saliency maps. The DVS model performed significantly better than the text only model as measured by six of the seven metrics (all *t*s > 2.04, all *p*s < 0.02). The only exception was the EMD metric ($t(391)$ = 0.65, p = 0.26). In this case, the scores for the text only and DVS models were nearly identical.

Figure 3. (A) An image from the MASSVIS set overlaid with fixation data and saliency maps produced by the original Itti (B), modified Itti (C), text saliency (D), and DVS (E) saliency models, with the DVS map overlaid on the original image in (F).

Table 1. Percentage Improvement for the DVS Model Relative to the Itti, BMS, eDN, and Text-Only Models.

		Itti	BMS	eDN	Text
Location Metrics	**AUC-J**	9%	11%	24%	2%
	AUC-B	9%	12%	22%	5%
	sAUC	9%	11%	21%	4%
Distribution Metrics	**SIM**	9%	14%	18%	15%
	EMD	18%	21%	26%	-1%
	CC	41%	70%	133%	5%
	KL	20%	37%	33%	--
Value Metric	**NSS**	55%	82%	176%	2%



.

Figure 4. Effect size, using Glass's delta, of the improvement due to using the DVS model for all eight metrics.

4 Testing the DVS Model's Performance

While the DVS model outperformed the Itti model in our initial assessment, there are several factors that limit our ability to generalize these findings. First, the MASSVIS data were collected in the context of a memory study, which might bias participants to focus more on the text in the visualizations. In addition, participants in the MASSVIS study viewed the images for 10 seconds, which is a longer duration than is typically used for comparing fixation data to saliency maps. For example, the widely-used eye tracking data sets provided by the MIT Saliency Benchmark had images that were presented for three seconds (MIT300) [27] or for five seconds (CAT2000) [4].

To get a broader understanding of the performance of the DVS model relative to existing saliency models, we used an additional data set to compare the performance of the DVS, Itti, BMS, and eDN models. This data set [37] consisted of eye tracking data collected from 30 participants who viewed four types of stimuli. As in the CAT2000 dataset, the participants viewed each stimulus for five seconds under free viewing conditions. The stimuli were presented in four counterbalanced blocks. One block contained 35 data visualizations from the MASSVIS dataset. Another contained 27 newly-generated, simple data visualizations that contained relatively little text. This set contained three visualizations of each of the following types: bar charts, box plots, bubble plots, column charts, line plots, parallel coordinates plots, pie charts, scatter plots, and violin charts. The other two blocks contained stimuli from the CAT2000 dataset [4] that were selected for their visualization-like properties. One block contained 30 line drawings and the other contained 16 images of fractals. These materials were chosen because they have already been incorporated into assessments of visual saliency models, yet like data visualizations, they have visual properties that differ from those of natural scenes. The line drawings have the same overall spatial layouts as natural scenes, but no colors and many fine contours that may be lost when the images are smoothed and resized by the saliency models. The fractals have very different spatial properties and color palettes than natural scenes, with vivid colors and shapes that fill the entire frame. Like data visualizations, they are abstract and computer-generated.

For each subset of stimuli, we assessed the match between the human fixation data collected by Matzen and colleagues [37] to the saliency maps produced by the DVS, Itti, BMS and eDN models using the eight MIT Benchmark metrics. In addition, as a point of reference, we compared the fixation data across experiments. For the MASSVIS stimuli, fixations were compared across the Matzen and colleagues [37] dataset and the original MASSVIS study [5]. For the fractal and line drawing stimuli, the fixation data was compared to the MIT Saliency Benchmark fixation data [4,7]. Although different groups of participants viewed the stimuli in the various experiments, and in the case of the MASSVIS data, the participants were performing a different task, we would expect to see the highest scores on the eight metrics when comparing one set of human fixations to another. If the models can accurately predict where viewers will look in data visualizations, their performance should approach the level of agreement between the two sets of fixation maps.

The results of the analysis for the line drawing stimuli are shown in Table S1 in the Supplemental Materials. These stimuli are most similar to natural scenes in terms of their spatial properties. As expected, the comparison between the two sets of fixation data had the best similarity scores for most of the metrics (six of the eight). When comparing the performance of the four models against the Matzen and colleagues [37] fixation data, the eDN model had the best scores for four of the eight metrics, the Itti model had the best scores on three of the metrics, and the DVS model had the best score on one metric, the sAUC.

The results of the analysis for the fractal stimuli are shown in Table S2 in the Supplemental Materials. These stimuli are somewhat of an intermediate point between natural scenes and data visualizations. They are computer generated and do not have naturalistic colors or spatial layouts, yet they do not contain text and their visual elements are not intended to convey specific information to the viewer. For these stimuli, the comparison of the two sets of fixation data had the best similarity scores for all eight metrics. When comparing the models to the fixation data, the eDN model had the best scores for six metrics and the DVS model had the best scores for two of the metrics.

The results of the analysis for the simple data visualizations are shown in Table S3 in the Supplemental Materials. When the four sets of saliency maps were compared to the fixation data, the DVS model had the best scores for seven of the eight metrics. The Itti model had the best score on the AUC-Borji metric.

The results of the analysis for the MASSVIS stimuli are shown in Table S4 in the Supplemental Materials. Once again, the comparison of the two sets of fixation data led to the best similarity scores for all eight metrics. When comparing the models to the fixation data, the DVS model had the best scores for all eight metrics.

To test whether or not the DVS model performed significantly better than the Itti, BMS and eDN models for data visualizations, the two sets of visualizations were combined. One-way ANOVAs were conducted for each of the eight metrics. These ANOVAs showed that there was a significant difference in performance across models for all eight metrics (all *F*s > 22.37, all *p*s < 0.001). Post-hoc t-tests showed that the DVS model's scores were better than the other models' scores for seven of the eight metrics (all *t*s > 3.74, all *p*s < 0.001). The exception was the AUC-Borji metric. According to this metric, the DVS model performed significantly better than the BMS ($t(61) = 6.50$, $p < 0.001$) and eDN ($t(61) = 9.34$, $p < 0.001$) models, but not the Itti model ($t(61) = 1.20$, $p = 0.12$).

4.1 Discussion

Our comparison of the Data Visualization Saliency model to the Itti, BMS, and eDN models found that the eDN model was generally the highest performer for line drawings, images that are somewhat abstract, but that share the spatial properties of natural scenes. This is consistent with the eDN model's overall high performance on the MIT Saliency Benchmark, the source from which the line drawing stimuli were taken. Similarly, the eDN model was also the best performer for fractal stimuli, which were also drawn from the MIT Saliency Benchmark set. We observed that the eDN model tends to produce saliency maps with a pronounced center weighting. This aligns well to the fixation maps for the fractal stimuli, where participants tended to fixate most on the center of the images.

For the line drawing and fractal stimuli, the DVS model's performance was typically similar to, or slightly better than, that of the Itti model, the model on which it is based. This indicates that our changes to the Itti model's color maps and the addition of the text saliency maps does not hinder the model's performance on stimuli that are not data visualizations. We anticipate that this would be true for images of natural scenes as well. The improved color map provides small improvements to performance, while the text saliency map contains only zero values in a scene that has no text, so it does not impact the final DVS map for such scenes.

Since our focus is on developing a saliency model that can be used as an evaluation tool for data visualizations, those stimuli provide the most important test of the model's performance. Our test set included two types of data visualization stimuli: simple visualizations that contained minimal text, no contextual information, and no "chart junk," and in-the-wild visualizations culled from publications, which typically contained explanatory text, source information, and graphical elements chosen for aesthetic or branding reasons. For the simpler data visualizations, the DVS model had the best performance according to seven of the eight metrics, and for the more complex visualizations, it had the best scores for all eight metrics. These results show that modifying the color map of the Itti model and adding a new visual feature (text saliency) led to significantly better performance on data visualizations.

For the MASSVIS stimuli, we were able to compare fixation data recorded from two different populations of participants in two different experimental contexts [5,37]. This comparison is in some sense a benchmark for model performance. If the models can accurately predict human fixations, their performance should approach the level of similarity obtained by comparing two sets of fixation data. The DVS model's scores were the closest to the scores for the fixation-to-fixation comparison for all eight metrics, and for the sAUC and KL metrics, paired t-tests showed that there was not a significant difference between the two scores ($t(34) = 0.01$ for sAUC, $t(34) = 0.04$ for KL).

5 Applying the DVS Model

Our results indicate that, of the models tested, the saliency maps produced by the DVS model were the best match to maps of human fixations, approaching the level of fixation-to-fixation comparisons in some cases. This suggests that the DVS saliency maps provide a reasonable approximation of which regions of a visualization are most likely to draw the viewer's attention.

As described above, this provides a useful evaluation metric for visualization designers. Ideally, the most important information in a visualization will also be highly salient [26,38]. Jänicke and Chen [26] illustrated this approach by using the Itti model as an evaluation tool. They compared saliency maps generated by the Itti model to a "relevancy map" defined by the visualization designer. They suggest that this comparison can be used to evaluate different visualization techniques or candidate visualizations in order to choose the one that most effectively highlights the important information.

The DVS model represents an improvement over the Itti model, but it can be used in a similar manner to evaluate visualizations. For example, the DVS saliency map in Figure 3 shows that the viewer's attention is most likely to be drawn to the text, the dark blue bars, and the tops of the light blue bars upon his or her initial viewing of the visualization. However, suppose that the visualization designer knows that the data represented by the line graphs is particularly important. The DVS saliency map provides a quick and easy way to assess whether or not this visualization will draw attention that data. In this example, the line graphs are not very salient, so the match between the importance of the data (i.e., top-down goals) and its salience (i.e., bottom-up attention) is poor. Armed with this information, the designer can try other variants of the visualization or other visualization techniques in order to select one that makes the most important information more salient.

The simplest way to evaluate a visualization using a saliency model is to take a qualitative approach. A designer can generate saliency maps for a set of visualizations and compare them visually, identifying the options that have a good distribution of saliency (as defined by the designer's goals). However, the saliency maps can also be used in a quantitative fashion. As suggested by Jänicke and Chen [26], designers could define a relevancy map and assess the match between the relevancy and the saliency maps. This assessment could be done categorically, as in their paper, or it could be done using one or more of the eight metrics that are commonly used to assess saliency maps. If only one is used, we propose that the value-based NSS metric would be the most appropriate for this type of comparison. If the designer assigns a relevancy value to each region of a visualization, the NSS metric can be used to assess the match between the relevancy values and the saliency values at each location. One prior study [23] has used the NSS metric to compare fixation data to important features in 2D flow visualizations, so there is some precedent for using this particular metric in the context of evaluating visualization techniques.

Another approach to quantitative assessment is to define regions of interest that outline the most important features in the data. After generating a saliency map, a designer could assess what percentage of the saliency falls within the regions of interest. This provides a simple numerical assessment of the match between the importance of the data and its saliency. To aid in evaluation, we have implemented this feature in the DVS model. A user can input the coordinates of a polygon describing a region of interest, and the model will provide the percentage of visual saliency, normalized for overall area, that falls within that region.

6 General Discussion

Visual saliency models have been the focus of a great deal of research in the cognitive science and computer vision communities because mimicking human visual attention has numerous applications, including image compression, image segmentation, object recognition, visual tracking, and image quality assessment [38,45,49]. Visual saliency maps could also play a role in evaluating data visualizations by allowing designers to determine whether or not a particular visualization draws the viewer's attention as intended. Since saliency models are inspired by the properties of the human visual system, the same system that is used to convey information in data visualizations, these models have the potential to serve as a simple and general evaluation tool.

While visual saliency models have a great deal of potential as an evaluation metric, prior evaluations have shown that existing saliency models consistently underperform on data visualizations, often failing altogether [18]. The models that perform best with natural scenes perform worst on data visualizations, and vice versa. Through assessments of three saliency models that generally perform well for natural scenes, we found that the spatial scales and visual features used by the existing saliency models are inadequate for data visualizations. Two particularly problematic areas were color models and text. The existing models perform operations using color spaces that do not correspond well to human perception of color. And while text draws a great deal of human attention, it is typically missed by saliency models due to its small spatial extent and high-frequency variation. Color and text are both very important features of data visualizations, chosen by designers to convey specific information to viewers. Thus, we chose to focus on these two areas in order to develop a saliency model that makes more accurate predictions of where viewers look in data visualizations.

We based the Data Visualization Saliency (DVS) model on the Itti model, which performed better than other existing saliency models on data visualizations. We modified the Itti model to use the CIE LAB color space, which is more representative of human color perception, and added a model of text saliency. We used a linear combination to incorporate the text saliency maps into the modified Itti model, and optimized the weighting of each component by testing the model against the stimuli in the MASSVIS dataset. To assess the performance of the final, weighted model, we compared its performance to the original Itti, BMS and eDN models using a set of fixation data obtained from participants viewing line drawings, fractals, and data visualizations [37]. We found that the DVS model's performance was comparable to the original Itti model's performance on the line drawing and fractal stimuli, and that it performed significantly better than the other models for data visualizations.

We suggest that the resulting model could be a simple and useful evaluation tool, which visualization designers can use to compare candidate designs in either a qualitative or quantitative manner. This approach is broadly applicable, but it may be particularly relevant to the evaluation of emphasis effects. There are numerous techniques that have been developed to emphasize subsets of the data in a visualization (see [19] for a review and evaluation framework). Hall and colleagues [19] frame emphasis effects in terms of visual prominence, which is another way of describing visual salience. They discuss intrinsic prominence, driven by the initial process of creating a visual mapping for data, and extrinsic emphasis effects, such as zooming and highlighting, that are used to enhance the prominence of selected features. Saliency maps could be used to evaluate both types of effects and to determine when one type of emphasis overrides the other. An evaluation based on visual saliency is particularly suited to assessing emphasis effects, since many of the features that are commonly used for emphasis (e.g., changes in color or size) are the same features that are used by saliency models.

Evaluations using visual saliency maps are complementary to other evaluation techniques, such as eye tracking. Eye tracking is a useful evaluation tool in its own right, and has been growing in popularity [13,15,16,29,44]. In our prior work with scene-like visualizations, we showed that eye tracking and saliency maps could be used in combination to assess the importance of features in the data and to understand the impact of users' expertise on their attention to those features. This provides information about how the visualization could be modified to better support the users' needs [38]. However, while eye tracking can be very informative, these studies can also be very time consuming and complex. Saliency maps provide a prediction of where users are likely to look without the need for eye tracking, and for many evaluation contexts, this may be sufficient.

6.1 Limitations and Future Directions

Although this model has the potential to be a simple and generalizable evaluation metric, there are several limitations to this approach. One important limitation is that the DVS model currently applies only to static images. This is a limitation both because interactions are a key component of many visualizations and because motion is a visual feature that typically captures human attention. In its current implementation, the DVS model can be applied to still images representing different phases of an interactive process, but it cannot capture the interactive component itself. In future work, motion detection algorithms could be incorporated into the model, enabling it to predict which parts of a dynamic scene will draw the viewer's attention most strongly. This would improve the model both in terms of its representation of human visual processing and in terms of its utility as an evaluation tool.

Another limitation is that the current implementation of the model does not change the spatial scales used by the Itti model, although these can also be problematic when applied to visualizations. The model resizes and smooths images, resulting in the loss of fine-grained details that are often very important in data visualizations. In future work, we plan to address these issues by allowing larger input images (limiting the need for resizing) and exploring the effects of changing the scales at which multiresolution differences are calculated.

A limitation of saliency models in general is that they focus on bottom-up visual attention. Bottom-up attention is only part of the picture, and top-down visual attention, driven by the viewer's task, goals, and prior experience, is also of tremendous importance in determining where a person will look in an image or a visualization [22,38,47]. Viewers with different goals may look at completely different parts of the same visualization. The DVS model incorporates one aspect of top-down attention by incorporating attention to text. Small regions of text may not be very salient from a bottom-up perspective, but people look at these regions because they expect them to convey meaningful information. In the future, additional feature detectors could be incorporated into the model to capture common graphical codes that convey semantic information in data visualizations [46], as these would also have high importance from the perspective of top-down attention. The eight evaluation metrics could be used to assess how the performance of the model changes with the addition of each feature.

On the other hand, the addition of more top-down features could quickly reduce the generalizability of the model. Text is unique in some sense because all literate people have extensive experience with processing text, to the point where it becomes automatic and involuntary [28,33,35]. That is not necessarily the case for other features that are used in visualizations. This could lead to differences between users with different levels of experience with the visualization technique or with the domain.

An alternate approach may be to incorporate Gestalt-based features into the model, since many visualization techniques are rooted in Gestalt psychology [46]. Like text comprehension, Gestalt principles reflect general cognitive processes that are not dependent on knowledge of any particular domain. The BMS saliency model relies on the Gestalt principle of figure-ground segregation to identify figures within an image [18,48], so incorporating Gestalt principles into a saliency model is certainly feasible. The BMS model does not perform well for visualizations [18], indicating that this principle alone is not sufficient for our purposes. However, it may be possible to use a similar approach to implement Gestalt-based features within the DVS model. The combination of the modified Itti maps, text saliency maps, and Gestalt-based maps could further improve the model's performance. This is an area that we would like to explore in future research.

Visualizations serve a variety of functions and support a vast range of tasks, so there is an enormous range of factors that might influence the viewer's top-down, goal-oriented processing. The wide range of roles for visualizations is part of what makes evaluation difficult in the first place! Saliency models cannot solve this problem, even with the addition of more features that are inspired by top-down attention. However, despite their imperfections, they can still be a useful tool in a designer's evaluation tool kit. If a designer has a sense of what information is most important from a top-down perspective, she can then assess the visual saliency of her design to determine whether or not the most important features are also salient from a bottom-up perspective. This provides a simple and rapid assessment that can be used in a quantitative or qualitative fashion to inform the visualization's design.

Acknowledgements

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

References

[1] J Aloimonos, "Purposive and qualitative active vision," *Proc. 10th International Conference on Pattern Recognition,* pp. 346-360, 1990.

[2] J Atkinson, "The Developing Visual Brain" Oxford University Press, Oxford, UK., 2002.

[3] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 35, pp.185-207, 2013.

[4] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *CVPR 2015 Workshop on the Future of Datasets*. arXiv preprint arXiv:1505.03581. 2015.

[5] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M Bainbridge, C.S Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond Memorability: Visualization Recognition and Recall," *IEEE Trans. Visualization and Computer Graphics*, vol. 22, pp.519-528, 2016.

[6] M. Borkin, Z. Bylinskii, G. Krzysztof, N. Kim, A. Oliva, and H. Pfister, "Massachusetts (Massive) Visualization Dataset," http://massvis.mit.edu. 2017.

[7] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, A. Torralba, "MIT Saliency Benchmark," http://saliency.mit.edu. 2017.

[8] Z. Bylinskii, T. Judd, A. Oliva, Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models," *arXiv preprint arXiv:1604.03605*, 2016.

[9] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, pp. 679-698, 1986.

[10] S. Carpendale, "Evaluating Information Visualizations" *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. Stasko, J.-D. Fekete, C. North (Eds.), Springer, pp. 19-45, 2008.

[11] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions." *18th IEEE International Conference on Image Processing (ICIP),* pp. 2609-2612. IEEE, September 2011.

[12] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology*, vol. 14, pp. R850-R852, 2004.

[13] R. Etemadpour, B. Olk, and L. Linsen, "Eye-tracking investigation during visual analysis of projected multidimensional data with 2D scatterplots," *International Conference on Information Visualization Theory and Applications (IVAPP),* pp. 233-246. IEEE, 2014.

[14] M. D. Fairchild and R. S. Berns, "Image color-appearance specification through extension of CIELAB" Color Research & Application, vol. 18, pp. 178-190, 1993.

J. H. Goldberg and J. Helfman, "Comparing information graphics: A critical look at eye tracking," *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization,* pp. 71-78. ACM, April 2010.

J. H. Goldberg and J. Helfman, "Eye tracking for visualization evaluation: Reading values on linear versus radial graphs," *Information visualization*, vol. 10, 182-195. 2011.

[15] A. Gonzalez, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Text location in complex images," *21st International Conference on Pattern Recognition (ICPR),* pp. 617-620. IEEE, November 2012.

[16] M. J. Haass, A. T. Wilson, L. E. Matzen and K. M.& Divis, "Modeling Human Comprehension of Data Visualizations," *International Conference on Virtual, Augmented and Mixed Reality*, pp. 125-134. Springer International Publishing, July 2016.

[17] K. W. Hall, C. Perin, P. G. Kusalik, C. Gutwin, and S. Carpendale, "Formalizing emphasis in information visualization," *Computer Graphics Forum*, vol. 35, pp. 717-737. 2016.

[18] J. Harel. "A saliency implantation in MATLAB" http://www.vision.caltech.edu/~harel/share/gbvs.php. 2017.

[19] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proceedings of Neural Information Processing Systems (NIPS),* pp. 545-552, 2006.

[20] J. M. Henderson, J. R. Brockmole, M.S. Castelhano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes" *Eye Movements: A Window on Mind and Brain*, pp.537-562, 2007.

H. Ho, I. Yeh, Y. Lai, W. Lin, and F. Cherng, "Evaluating 2D flow visualization using eye tracking," *Computer Graphics Forum*, vol. 34, pp. 501-510. 2015

[21] L Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience, vol. 2*, pp. 194-203, 2001.

[22] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. *11*, pp. 1254-1259, 1998.

[23] H. Jänicke and M. Chen, "A salience-based quality metric for visualization," *Computer Graphics Forum*, vol. 29, pp. 1183-1192. Blackwell Publishing Ltd., 2010.

[24] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations" MIT Technical Report, 2012.

[25] D. Kahneman and D. Chajczyk, "Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 9, pp. 497, 1983.

K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf, "Evaluating visual analytics with eye tracking," *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 61-69. ACM, 2014.

[26] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios" *IEEE Trans. Visualization and Computer Graphics*, vol. 18*,* pp. 1520– 1536, 2012.

[27] G. E. Legge, S. J. Ahn, T.S. Klitz, and A. Luebker, "Psychophysics of reading—XVI. The visual span in normal and low vision" *Vision Research,* vol. 37*,* pp. 1999-2010, 1997.

[28] Y. Li and H. Lu, "Scene text detection via stroke width," *21st International Conference on Pattern Recognition (ICPR),* pp. 681-684. IEEE, November 2012.

[29] G. D. Logan, "Automaticity and reading: Perspectives from the instance theory of automatization" *Reading & Writing Quarterly: Overcoming Learning Difficulties,* vol. 13, pp. 123-146, 1997.

[30] S. Lu, T. Chen, S. Tian, J. H Lim and C. L Tan "Scene text extraction based on edges and support vector regression" *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, pp. 125-135, 2015.

[31] C. M MacLeod, "Half a century of research on the Stroop effect: An integrative review" *Psychological bulletin,* vol. 109, p. 163, 1991.

[32] J. Matas, O. Chum, M. Urban & T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, pp. 761-767, 2004.

L. E. Matzen, M. J. Haass, K. M. Divis and M.C. Stites "Patterns of attention: How data visualizations are read," *Augmented Cognition. Enhancing Cognition and Behavior in Complex Human Environments*, D. D. Schmorrow and C. M. Fidopiastis, eds., pp. 176-191. Springer, 2017.

[33] L. E Matzen, M. J. Haass, J. Tran, and L. A. McNamara, "Using eye tracking metrics and visual saliency maps to assess image utility," *Proc. Human Vision and Electronic Imaging (HVEI) XXI*, 2016.

[34] E. N. Merieb and K. Hoehn, "Human Anatomy & Physiology 7th Edition," Pearson International Edition, 2007.

[35] L. Neumann and J. Matas, "Real-time scene text localization and recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3538-3545. IEEE, June 2012.

[36] T. Ogawa and H. Komatsu, "Target selection in area V4 during a multidimensional visual search task" *Journal of Neuroscience*, vol. 24, pp. 6371- 6382, 2004.

[37] N. Pinto and D.D. Cox, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition" *IEEE Automatic Face and Gesture Recognition*, 2011.

[38] Y. Pinto, A. R. van der Leij, I.G. Sligte, V. A. F. Lamme, and H. S. Scholte, "Bottom-up and top-down attention are independent," *Journal of Vision,* vol. 13, pp. 1-14, 2013.

B. Strobel, S. Saß, M. A. Lindner, and O. Köller, "Do graph readers prefer the graph type most suited to a given task? Insights from eye tracking," *Journal of Eye Movement Research*, vol. 9, pp. 1-15. 2016.

[39] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798-2805, 2014.

Ware, Colin. *Information visualization: perception for design*. Elsevier, 2012.

[40] A. Yarbus, *Eye Movements and Vision*. New York City: Plenum Press, 1967.

[41] J. Zhang and S. Sclaroff "Exploiting surroundedness for saliency detection: A boolean map approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[42] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *Journal of Vision,* vol. 8, pp. 32-32, 2008.

DISTRIBUTION

1	MS0932	Michael Haass	9365
1	MS1326	Andrew Wilson	1461
1	MS1327	Ron Oldfield	1461
1	MS1327	Phil Bennett	1463
1	MS1327	Kristin Divis	1463
1	MS1327	Laura Matzen	1463
1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911

Sandia National Laboratories