Graduate Theses, Dissertations, and Problem Reports

2021

Investigating the Role of Attention and Memory in Visual Exploration

Jacob E. Suffridge
jesuffridge@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Cognitive Neuroscience Commons, and the Computational Neuroscience Commons

Recommended Citation
https://researchrepository.wvu.edu/etd/8311
Investigating the Role of Attention and Memory in Visual Exploration

Jacob E. Suffridge

Thesis submitted to the Benjamin M. Statler College of Engineering and Mineral Resources at West Virginia University in partial fulfillment of the requirements for the degree of

Master of Science in Biomedical Engineering

Shuo Wang, Ph.D., Chair
Jessica Allen, Ph.D.
Xin Li, Ph.D.

Department of Chemical and Biomedical Engineering

Morgantown, West Virginia
2021

Keywords: Free viewing, Eye tracking, Saliency Modeling, Memory, Visual Exploration

Copyright 2021 Jacob E. Suffridge
Abstract

Investigating the Role of Attention and Memory in Visual Exploration

Jacob E. Suffridge

This research investigates the role that attention and memory plays in visual exploration through a comprehensive analysis of eye movement and behavioral data, alongside incorporation of a computational saliency model. The purpose of this project is to quantify differences in visual attention over repeated viewings of natural scene images by employing a free viewing task that contains a memory component. In Chapter 2, we explore the task’s behavioral data showing that participants generally memorize our images well before we investigate the effect of numerous object and individual feature inclusion. In Chapter 3, we develop four primary methods to quantify visual exploration of our images and implement these methods to determine if differences exist between new and repeated images but also between remembered and forgotten images. We found that these methods showed more intensely focused visual attention in images that had been repeatedly viewed. In Chapter 4, we introduce a computation saliency model adapted from previous works (Wang et al., 2015) and discuss the extensive tuning needed to achieve optimal performance. Further, we explore the differences in saliency weights for new vs repeated images and also 1st vs 2nd vs 3rd image viewings, showing multiple significant differences were observed. In Chapter 5, we incorporate additional methods to determine if our natural scene images were viewed consistently. This included examining fixations in the order of occurrence but also total image consistency. The latter suggested that initial viewings are more similar to tertiary viewing than to secondary viewings albeit this trend is not observed as strongly in images that contain certain semantic features. In Chapter 6, we replicate our previously discussed methods on preliminary data taken from participants with ASD to explore difference in visual attention when compared against controls. Finally in Chapter 7, we discuss key results and conclusions developed by this research before commenting on ways to further alter and enhance the included analyses to provide a more comprehensive analysis of visual exploration.
Acknowledgements

Thank you to my advisor, Dr. Shuo Wang, for your support, guidance and patience. I have gained a vast wealth of knowledge ever since we began working together during the senior year of my undergrad. I am extremely grateful that you allowed me to become a part of your lab, learn new research techniques and encounter experiences that I otherwise would not have had. I especially want to thank you for your extended support and patience during the hardships of the 2020 pandemic. I appreciate all that you have and continue to do for me.

I would also like to thank all my friends, colleagues and especially my committee, Dr. Jessica Allen and Dr. Xin Li, for all their help, advice and direction over the years. Also, I want to express gratitude to Michael, Fosco and Nathan from EPFL for allowing me to contribute and learn from the surgeries and data collection associated with their DBS protocol.

Finally, I would like to thank my family, especially my parents and both of my brothers, for all their help over the last several years. Know that your continuous support and encouragement as not gone unnoticed or unappreciated.
# Table of Contents

Abstract...................................................................................................................................................... ii

Acknowledgements ................................................................................................................................. iii

Table of Contents....................................................................................................................................... iv

Table of Figures ........................................................................................................................................ vi

List of Tables: ........................................................................................................................................... ix

Chapter 1: Background ............................................................................................................................... 1
  1.1: *Overview / Motivation* .................................................................................................................... 1
  1.2: *Free Viewing Memory Task* ........................................................................................................ 2
  1.3: *Image Set* ..................................................................................................................................... 3
  1.4: *Control Data* .................................................................................................................................. 4

Chapter 2: Patterns of Behavior and Memory ......................................................................................... 5
  2.1: *Overview* ...................................................................................................................................... 5
  2.2: *Methods* ...................................................................................................................................... 6
  2.2: *Results / Discussion* ...................................................................................................................... 7

Chapter 3: Fixation and Gaze Exploration ............................................................................................... 16
  3.1: *Overview* .................................................................................................................................... 16
  3.2: *Methods* ..................................................................................................................................... 17
  3.3: *Results / Discussion* .................................................................................................................... 18

Chapter 4: Saliency Modeling .................................................................................................................. 21
  4.1: *Overview* .................................................................................................................................... 21
  4.2: *Methods* ..................................................................................................................................... 21
  4.3: *Results / Discussion* .................................................................................................................... 27

Chapter 5: Viewing Consistency ............................................................................................................... 32
  5.1: *Overview* .................................................................................................................................... 32
  5.2: *Methods* ..................................................................................................................................... 32
  5.3: *Results / Discussion* .................................................................................................................... 36
    5.3.1 Viewing Consistency .................................................................................................................... 36
    5.3.2 Correlation Analysis .................................................................................................................... 41

Chapter 6: Comparison of Visual Attention in ASD ............................................................................... 45
  6.1: *Overview* .................................................................................................................................... 45
# Table of Figures

Figure 1.1 Button box displayed on screen between image trials .................................................. 2
Figure 2.1 Total image set behavior: Distribution of trials (A), Average Fixation Duration (B),
Button-press reaction time (C), Average Confidence Response (D) .................................................. 7
Figure 2.2 More Objects vs Less Objects: Distribution of trials (A), Average Fixation Duration
(B), Button-press reaction time (C), Average Confidence Response (D) ........................................... 8
Figure 2.3 Faces vs No Faces: Distribution of trials (A), Average Fixation Duration (B), Button-
press reaction time (C), Average Confidence Response (D) .......................................................... 9
Figure 2.4 Emotion vs No Emotion: Distribution of trials (A), Average Fixation Duration (B),
Button-press reaction time (C), Average Confidence Response (D) .................................................. 9
Figure 2.5 Gazed vs No Gazed and Operability vs No Operability: Distribution of trials (A),
Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D) ....... 10
Figure 2.6 Sound vs No Sound: Distribution of trials (A), Average Fixation Duration (B), Button-
press reaction time (C), Average Confidence Response (D) .......................................................... 11
Figure 2.8 Taste vs No Taste and Touch vs No Touch: Distribution of trials (A), Average
Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D) ......... 12
Figure 2.7 Smell vs No Smell: Distribution of trials (A), Average Fixation Duration (B), Button-
press reaction time (C), Average Confidence Response (D) .......................................................... 12
Figure 2.9 Watchability vs No Watchability: Distribution of trials (A), Average Fixation Duration
(B), Button-press reaction time (C), Average Confidence Response (D) ........................................... 13
Figure 2.10 Motion vs No Motion: Distribution of trials (A), Average Fixation Duration (B),
Button-press reaction time (C), Average Confidence Response (D) .................................................. 14
Figure 2.11 Text vs No Text: Distribution of trials (A), Average Fixation Duration (B), Button-
press reaction time (C), Average Confidence Response (D) .......................................................... 14
Figure 2.12 Touched vs No Touched: Distribution of trials (A), Average Fixation Duration (B),
Button-press reaction time (C), Average Confidence Response (D) .................................................. 15
Figure 3.1 Natural scene image viewed for the first time (left) and viewed a second time (right).
Yellow circles illustrate subject fixations .......................................................................................... 16
Figure 3.2 Natural scene image viewed for the first time (left) and viewed a second time (right).
Subject fixations and fixation lines displayed in yellow, dispersion area is shown as the region
contained within the red boundary ................................................................................................. 17
Figure 3.3 Natural scene image viewed for the first time (left) and viewed a second time (right).
Subject fixations displayed in yellow, center distances are depicted in magenta ............................. 18
Figure 3.4 New vs repeated Images: Number of Fixations (A), Average Fixation Duration (B), Fixation Dispersion (C) and Center Bias ................................................................. 19
Figure 3.5 Remembered vs Forgotten Images: Number of Fixations (A), Average Fixation Duration (B) Fixation Dispersion (C) and Center Bias ................................................................. 19
Figure 4.1 Preprocessing of fixation map (left) and isolated fixations (right) .................................................. 22
Figure 4.2 Convolution of identical fixation maps with patch size 30 and std 15 (left) vs patch size 30 and std 10 (mid) vs patch size 10 and std 15 (right). .................................................. 23
Figure 4.3 Difference in saliency weights (left) generated by Test 7 and Test 14 for new vs repeated images alongside relative difference in weights (right) ............................................. 26
Figure 4.4 Average (left) vs Best Performing weights (right) for Semantic + Background and Center Features. ............................................................................................................. 28
Figure 4.5 Relative Difference in average weights vs the best performing weights for 1st, 2nd and 3rd viewings .................................................................................................................. 28
Figure 4.6 Saliency weights generated by the most optimal new vs repeated model for all features (left) and all features excluding center (right). ........................................................ 29
Figure 4.7 Saliency weights generated by the best performing model on 1st, 2nd and 3rd viewings. ......................................................................................................................... 29
Figure 5.1 Ordered fixations from novel and Repeated image viewing with semantic objects highlighted in yellow, novel fixations in green and repeated fixations in blue ............................................ 33
Figure 5.2 Ratio of trials where ordered fixations occupied the same feature categories in both novel and repeated images. ........................................................................................................ 36
Figure 5.3 Ratio of trials where ordered fixations occupied the same feature categories in both the novel and repeated viewings of images that contained the given semantic feature .......... 37
Figure 5.4 Viewing consistency ratios for all images disregarding fixation order .............................................. 38
Figure 5.5 Ratio of trials where fixations occupied the same feature categories in both the novel and repeated viewings of images that contained the given semantic feature, note this disregards fixation order. ............................................................................................................. 39
Figure 5.6 Average number of trials where a fixation occurred exclusively in the first or second viewing normalized by feature count. ........................................................................... 40
Figure 5.7 Average number of trials where a fixation occurred exclusively in the first or third viewing normalized by feature count. ........................................................................... 41
Figure 5.8 Average correlation coefficient between viewings .................................................................................. 42
Figure 5.9 Correlation analysis of fixation density maps with ordered vs shuffled labels for all images ................................................................................................................................. 43
Figure 5.10 Correlation Comparison between repeated viewings of images with greater than or less than average objects

Figure 6.1 Total image set behavior ASD: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D)

Figure 6.2 Comparison between Control and ASD behaviors: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C)

Figure 6.3 New and repeated Images for Control vs ASD: Number of Fixations (A), Average Fixation Duration (B), Fixation Dispersion (C) and Center Bias (D)

Figure 6.4 Remembered and Forgotten Images for Control vs ASD: Number of Fixations (A), Average Fixation Duration (B), Fixation Dispersion (C) and Center Bias (D)

Figure 6.5 Saliency weight comparison: Controls vs ASD for new Images

Figure 6.6 Saliency weight comparison: Controls vs ASD for Repeat Images

Figure 6.7 Control vs ASD: Ratio of trials where ordered fixations occupied the same feature categories in both Novel and Repeat Image

Figure 6.8 Control vs ASD: Viewing Consistency Ratios for all images disregarding fixation order
List of Tables:

Table 1: Compilation of parameter tests used to gauge the performance of our model on new vs repeated images. Note tests highlighted in green had the highest AUC and testing accuracy and tests weren’t necessarily completed in the order listed. .......... 24
Table 2: Compilation of parameter tests used to gauge model performance on 1st, 2nd and 3rd image viewings. Note tests highlighted in green had the highest AUC and testing accuracy and tests weren’t necessarily completed in the order listed .......... 25
Table 3: Tabulated differences observed in Figure 4.7 for comparison between 1st, 2nd and 3rd image viewings .......................................................... 31
Table 4 Breakdown of usable trials for each quantity of fixations selected ................. 34
Table 5 Mean correlation coefficient between all semantic subsets and their corresponding No-feature subset ........................................................................ 44
Table 6 Comparison differences between ASD and control saliency weights tabulated from Figures 6.5 and 6.6 ........................................................................ 50
Chapter 1: Background

1.1: Overview / Motivation

The study of human attention has become increasingly popular due to its important role in brain functionality and cognitive neuroscience. Investigation of visual attention by Ramanathan et al. has shown visual attention is not simply subjective but instead is directed toward salient objects, seen by the consistent presence of fixations clustered around salient image locations. Additionally, saliency maps have become quite essential to attention-based research because of their role in gaze shift guidance. The visual cortex of primates is believed to include a two-stage selection process: bottom-up stimulus features and top-down attentional modulation (Itti et al., 2000). According to Treue (2003), these processes combine to reveal relevant information that is eventually analyzed while neglecting less important information. For an image, a saliency map is a two-dimensional topographical space that encodes stimulus conspicuity at all points of the image (Itti et al., 2001). More simply, this means a map of locations in an image that people pay attention to and in turn can be monitored by fixations using an eye tracking system. Research as early as Itti et al., 2000 provides computation modeling techniques used to weigh the importance of different features using a bottom-up saliency approach with low-level features. More recent research shows the added benefits of incorporating high-level features into saliency models such as object and semantic level features (Xu et al., 2014; Wang et al., 2015; deHaas et al., 2019 and Shaojing Fan et al., 2018). Our computational model, described in detail in Chapter 4, examines a combination of low-level, object-level and semantic-level features when calculating saliency summary statistics such as saliency weights.

Another component of our research is to study fixations during visual exploration and memory tasks. This exploration is motivated by recent studies exploring neuronal activity in primates using saccadic or rapid movement of the eye between fixations. Studies such as Jutras et al. (2013), have shown that not only is there a link between eye movement and hippocampus activity but also that saccadic eye movement could modulate neuron activity by altering the phase reset of hippocampus theta waves.
Additionally, electrophysiological evidence has been provided that reports the amygdala may have a significant contribution with saccadic eye movement in primates (Gonzalez et al., 2008). Furthermore, reports from Zhao et al. (2012) have shown that saccadic eye movements are strictly dependent on attention as pre-saccadic shifts of attention contribute to neural remapping of visual receptive fields. Such findings prompt the need for additional research on the linkage between attention and memory in regard to human eye movement. Similar to Rutishauser et al. our Free Viewing Memory task pairs viewing images freely alongside a memory task. While they focus on analysis of single-unit neuron firing rates in the hippocampus and amygdala, we focus on a detailed, comprehensive analysis of the eye tracking data and an in-depth look at features viewed within our natural scene image set.

1.2: Free Viewing Memory Task

The primary experiment for my research was the Free Viewing Memory (FVM) task where participants viewed natural scene images while being monitored by a non-invasive infrared eye tracking system. Prior to starting the task, participants underwent several, simple preliminary exercises of gazing at specified dots arranged in a 3-by-3 grid on the stimulus screen. These exercises ensured that our eye tracking system was properly and accurately calibrated before each session. During each image trial, participants were asked to freely view the given image for 3 seconds to ensure adequate time for eye movement. After each image viewing, participants were given 3 seconds to report whether they remembered seeing the image before and their confidence. They input their selection via a standard keyboard using the numbers 1 through 6 (from left to right) to correspond with the on-screen button box seen in Figure 1.1.

![Figure 1.1 Button box displayed on screen between image trials](image-url)
Participants were asked to view images across two sessions: a learning session and a recognition session. During the learning session, participants viewed 100 new images alongside 50 repeated images for a total of 150 image presentations. In the recognition session, participants viewed 100 entirely new images (not presented in the learning session), 50 once-repeated images (only presented in recognition session) and 50 twice-repeated images (presented in both learning and recognition sessions) for a total of 200 images presented. All participants viewed a random shuffling of the same 150 images and 200 images in the learning and recognition sessions, respectively. Between the two sessions a standard 15-minute break was administered to give participants a short break. To keep the conditions of the task as controlled as possible, we asked participants to complete questionaries for Autism Quotient (AQ), Beck Depression Index (BDI) and Social Responsiveness Scale (SRS) during the break. The typical duration for the FVM task was approximately one hour, however the duration did vary based on response times. Our FVM task provided many avenues for exploration including participant behavior, detailed analysis of the eye tracking data and even an in-depth look at features viewed within our natural scene image set.

1.3: Image Set

The test images utilized in our free viewing task were natural scene images taken from the OSIE dataset which has been described in detail previously (Xu et al., 2014, Wang et al., 2015). Briefly, the dataset contains 700 images, which have been quantified according to three pixel-level attributes (color, intensity, and orientation), five object-level attributes (size, complexity, convexity, solidity, and eccentricity), and twelve semantic attributes (face, emotion, touched, gazed, motion, sound, smell, taste, touch, text, watchability, and operability) annotated on a total of 5551 segmented objects. Since there are a large number and variety of objects in natural scenes, to make the ground truth data least dependent on subjective judgments, several guidelines were followed for the segmentation, as described in Xu et al., 2014. Images contain multiple dominant objects in a scene. The twelve semantic attributes fall into four categories: (i) directly relating to humans (i.e., face, emotion, touched, gazed); (ii) objects with implied motion in the image; (iii) relating to other (non-visual) senses of humans (i.e., sound,
smell, taste, touch); and (iv) designed to attract attention or for interaction with humans (i.e., text, watchability, operability). Alongside these features we also implemented image center and background features in our analysis (detailed in Chapter 4). The FVM task utilized a 200-image subset of the 700-image OSIE dataset that was confirmed to have adequate representation for each of the previously discussed features.

1.4: Control Data

I recruited 22 neurotypical individuals to participate as controls in the FVM task with a breakdown of 12 males and 10 females, most of which were adults in graduate school or undergraduate students. Ultimately, data collected from 3 of the female participants was excluded due to significant data loss and/or corruption. 2 of the 3 were missing a large percentage of their eye tracking data and the other subject’s entire recognition session was lost. This resulted in 12 male and 7 female (mean age 22.63 ± 3.86) participants with usable eye tracking and behavioral data. This group of controls will be the subject of analysis in Chapters 2, 3, 4, 5 before being compared to participants with autism spectrum disorder (ASD) in Chapter 6.
Chapter 2: Patterns of Behavior and Memory

2.1: Overview

The FVM task provided a very rich dataset to analyze, first we were interested in the behavioral data generated by the memory portion of the task. Across both the learning and recognition sessions, subjects were asked to report whether they remembered seeing the presented image. The correctness of their responses produced four types of memory trials:

- **Hit**: remembered image — a repeated image correctly reported as “seen before”
- **Miss**: forgotten image — a repeated image incorrectly reported “not seen before”
- **Correct rejection**: a new image correctly reported as “not seen before”
- **False alarm**: a new image incorrectly reported as “seen before”

In addition to these trial types, we were able to extract the duration of subject fixations, their button-press response time and also their confidence in their selection. In this section, we are specifically interested in accessing our participants memory by analyzing how these four memory types were distributed across trials but also how the fixation duration, reaction time and confidence were affected by memory.

Another interesting component we explored was how changing the images effected our results. Our first round of analysis was performed on all 200 images in our subset. For following rounds, we began to further partition the 200 images into groups such that one group had all the images where a given feature was present and the other group had all images where that feature was not presented. This was particularly interesting because it gave us a method to test how feature inclusion effected memory, fixation duration, reaction time and confidence ratings. We proceeded to perform this comparison for all 12 semantic features discussed in Chapter 1. A final test was included in our subset comparisons to explore the effect of object presence in our images. Simply, we compared the metrics between images with more objects versus those containing less objects.
2.2: Methods

All calculations and statistics for this chapter and most proceeding chapters were performed in MATLAB. To access the participants memory, we look at the distribution of memory events. The structure of the data returned by the FVM task allows for a simple calculation of which images were remembered correctly and those that were not. Further splitting the correct and incorrect by whether or not the image was new or repeated give us the distribution of our four memory events. Due to the nature of the calculation, it should be noted that hit and miss will always add to 1 as will the sum of false alarm and correct rejection. The button-press reaction time is returned for each subject for each trial. To analyze the reaction time, we average across the trial times to get an average reaction time per person before averaging across subjects. This is also the method used to determine the average confidence except prior to being averaged, the confidence ratings are converted from the 1 to 6 down to a 1 to 3 scale where 1-sure, 2-less sure and 3-very unsure. Essentially, this just removes the yes or no component of their response which is no longer needed for the average confidence ratings. The final metric explored was the fixation duration. The fixations were extracted from the raw eye tracking data using the built-in software provide by the Eyelink 1000 system. The fixation duration was simply calculated as the time duration between saccades. The average fixation durations were then found similarly to the previous metric calculations. Note in all subsequent comparisons and figures, error bars are used to illustrate the standard error of the mean across subjects.

To construct the subsets that we want to compare, we first had to filter our images by feature. Luckily the OSIE dataset images were heavily documented, providing which features were included in each image. We then constructed two subsets, one with the feature included and one without, for each semantic feature. For the object comparison, we simply calculated the number of objects presented in each of our images to provide an average number of objects. To construct the two subsets for comparison, we took one group as all images with more than the average number of objects and the other as the images with less than the average number of objects.
Redoing our previously discussed analysis provided us with 13 sets of comparisons that we further discuss in the next section.

### 2.2: Results / Discussion

From the 38 usable sessions, fixation data was extracted from the captured eye tracking data and sorted into their corresponding memory events. We first show in **Figure 2.1** that participants can memorize images well with very high average hit and correct rejection rates. Note the vastly lower percentage of fixations for the incorrect response types of miss and false alarm. The fixation durations are also notably different between Hit and Miss (t-test, p = 0.005) and Correct Rejection and False Alarm (t-test, p = 0.025). Note for both the repeated (Hit and Miss) and new (Correct Rejection and False Alarm) trial types, the correct response has a higher average fixation duration when compared to the incorrect response. In particular, a higher fixation duration means that in correct response trials, participants gaze at fewer locations but fixate on those locations for longer.

![Figure 2.1 Total image set behavior: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (**p<0.01, *p<0.05, error bars denote SEM across subjects)](image)

No notable trends are revealed by the reaction time given all four memory trials had very similar average reaction times. The confidence also did not differ significantly across the four memory trials (t-test, p = 0.073 and p = 0.111), but it should be noted
that overall subjects show very high confidence in their selections. Average confidence ranged from $1.132 \pm 0.061$ to $1.261 \pm 0.072$ with confidence selection values taken as 1, 2 or 3. This observation implies that participants selected a confidence rating of very sure for the majority of trials.

Further, we repeat a similar analysis but instead examined the differences in images that contain and don’t contain semantic features and those with more or less than average objects. Surprisingly in the object comparison in Figure 2.2, we see very similar trends in trial distribution, fixation duration, reaction time and confidence. From this result, we concluded that the number of objects presented in the image did not have a particular effect on behavioral performance.

![Figure 2.2 More Objects vs Less Objects: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (Error bars denote SEM across subjects)](image)

Next, we explored the behavioral effect that each semantic feature has on performance. An example considering images with faces and images without faces is illustrated in Figure 2.3. While fixation durations, reaction times, and confidence did not differ in face images compared to those without, the distribution of trials for all four memory types were significantly different. Surprisingly the correct response types, hit and correct


rejection, had lower rates in images that contained the semantic face feature when compared to those without. This suggests that faces could actually decrease the likelihood that trials would be correctly remembered. Another semantic feature we
showcase in Figure 2.4 is emotion. When comparing images with emotion versus those without emotion features, we observed the opposite effect of the face image comparison. Correct response types had higher rates in images with emotional objects than those without, suggesting that emotion features could increase the likelihood of correctly remembering those images.

![Figure 2.5 Gazed vs No Gazed and Operability vs No Operability: Distribution of trials (A), Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (**p<0.05 *p<0.05, error bars denote SEM across subjects)](image)

Noteworthy differences were observed in the feature comparisons of the Gazed, Operability, Smell, Sound, Taste, Touch and Watchability semantic features. In the gazed comparison, only the correct rejection and false alarm rates displayed any significant difference. The correct rejection rate was slightly lower (false alarm rates were higher) in images without the gazed feature meaning inclusion of the gazed feature could correlated with an increase in correct rejection rates. A very similar trend is observed in the operability feature comparison which is also seen in Figure 2.5. Of all comparisons, only sound and smell comparisons showed any notable differences in the button-press reaction times. In the sound comparison of Figure 2.6, both hit and correction rejection displayed higher reaction times in images with sound objects than those without. The Figure 2.7 smell comparison shows, the hit reaction times trended
higher in images with smell object compared to those without. Note that correct response types, hit and correct rejection, rates were also significantly higher in both smell and sound comparisons. This suggests that smell and sound features increased the likelihood of images being correctly remembered but these higher correct response rates could also be explained simply as longer reaction time providing more time to consider each image.

Figure 2.6 Sound vs No Sound: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (**p<0.01, *p<0.05, error bars denote SEM across subjects)
Figure 2.7 Smell vs No Smell: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (**p<0.01, *p<0.05, error bars denote SEM across subjects)

Figure 2.8 Taste vs No Taste and Touch vs No Touch: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (*p<0.05, error bars denote SEM across subjects)
The Taste and Touch feature comparisons are both contained in Figure 2.8 due to the similarities in their differences. Hit rates were significantly higher in images with Taste and Touch features than those without their respective feature. This suggests that taste and Touch features increased the likelihood of new images being correctly processed (i.e., higher hit and lower miss rates). The opposite trend is observed in Figure 2.9 for the watchability feature. Lower hit rates are observed in the images containing watchability objects compared to those without these objects. This suggests that the watchability feature could decrease the likelihood of new images being correctly processed (i.e., lower hit and higher miss rates).

The remaining comparisons for the three semantic features: Motion, Text and Touched did not exhibit any significant differences. The figures showcasing these comparisons are included below in Figure 2.10-12 for completeness.
Figure 2.10 Motion vs No Motion: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (Error bars denote SEM across subjects)

Figure 2.11 Text vs No Text: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (Error bars denote SEM across subjects)
Figure 2.12 Touched vs No Touched: Distribution of trials (A), Average Fixation Duration (B), Button-press reaction time (C), Average Confidence Response (D). (Error bars denote SEM across subjects)
Chapter 3: Fixation and Gaze Exploration

3.1: Overview

In this section we shift our attention from analyzing behavior to take a comprehensive look at fixations to examine the link between memory and eye movement. Fixations can be simply defined as objects or areas of interest that capture an observer’s gaze. Using our eye tracking system and built-in software provided by Eyelink, we are able to capture and visualize these fixations as seen in Figure 3.1. This is a particularly interesting example because, while multiple objects are present in the image, only a select few objects attracted the gaze of the observer. Also, note the substantial differences in the visual exploration of the image in the first presentation compared to that of the second presentation. These inherent distinctions between viewings serve as our motivation to explore how memory and repeated viewings impact fixations and visual exploration of images. To quantify and study these effects, we devised three new metrics in addition to fixation duration. These metrics are the number of fixations, dispersion of fixations and the bias of fixations toward image center. The latter metric was inspired by the strong image center effect seen in those with ASD and will be explored in much further detail in Chapter 6.
3.2: Methods

We quantify visual exploration using four metrics: number of fixations, fixation duration, dispersion area of fixations and center bias of fixations. Fixations were extracted from the raw eye tracking data using software provided by the Eyelink 1000 system. The fixation duration was calculated in the same manner discussed in Section 2.2 by calculating the time duration between saccades. The first of our new metrics is a straightforward calculation of the average number of fixations observed. Like all other metrics discussed in this section, we average across images and then across subjects for statistical comparisons. The latter two metric introduced in this section require a bit more explanation. We first sought out to quantify visual exploration by introducing a metric to quantify the dispersion of fixations. We define the fixation dispersion as

the area, measured in square pixels, encompassed by a convex polygon whose edges are drawn from the outermost fixations. Note that this definition forces all fixations to be contained within or on the boundary of the given polygon as illustrated in Figure 3.2. Once this polygonal area is determined, the fixation dispersion is calculated directly as the percentage of the polygon area divided by the total image area. Note that all images across all chapters were presented to the stimulus screen at a constant 1600 by 1280 resolution. The final metric is the image center bias showcased in Figure 3.3. We define the center bias merely as the average distance fixations occur from the image center.

\[ \text{Fixation Dispersion} = \frac{\text{Area of Polygon}}{\text{Total Image Area}} \times 100 \% \]
Now that we have described our quantification metrics, we must also determine the groups we decided to test these metrics on. For current and following chapters, we classify each of our trials into four distinct event types:

- **New Image**: image seen for the first time
- **Repeated Image**: image that has been presented two or more times
- **Memorized**: image that has been correctly remembered
- **Forgotten**: image that has not been correctly remembered

We are especially interested in comparisons between new and repeated images (seen inherently in Figures 3.1, 3.2 and 3.3) but also for memorized vs forgotten trials.

### 3.3: Results / Discussion

The first comparison we explore here is new or novel images versus repeated images. Applying all of the metrics defined in the methods of section 3.2 section, provides us the results seen in Figure 3.4. The number of fixations decreases in repeated viewings ($t$-test, $p = 6.03e-06$) resulting in an increase in fixation duration ($t$-test, $p = 4.585e-4$) as seen in A and B of Figure 3.4. The inversely coupled relationship of the number and duration of fixations was to be expected, considering the viewing window was held at a constant 3 seconds. We suspect that the trend in the number of fixations and the fixation durations were due to participants displaying a more focused attention in additional viewings resulting in less visual exploration. Further analysis showed that
that the dispersion of fixations on repeated images (C in Figure 3.4) was significantly lower (t-test, p = 3.53e-05) than that of new images. Also, the average distance of the fixations compared to the image center (D in Figure 3.4) was significantly lower (t-test, p = 1.760e-3) for repeated images. Together these metrics support our conclusion that repeated viewings of the same image led to a more focused visual attention.

Figure 3.5 Remembered vs Forgotten Images: Number of Fixations (A), Average Fixation Duration (B) Fixation Dispersion (C) and Center Bias. (t-test, **p<0.01, error bars denote SEM across subjects)
The last comparison we investigate in this section is between memorized (or remembered) images versus those that were forgotten. Applying all of our previously defined metric to the sorted trials provides the results presented in Figure 3.5. Of our test criterion, only the number of fixations proved to differ significantly for remembered vs forgotten images. Forgotten images had fewer average fixations compared to remembered images (t-test, p = 0.003). While not all four measures differ significantly, you can see trends developing in the memorized versus forgotten that are very similar to that of new images versus repeated images. The forgotten images appear to behave like images that have never been seen before, suggesting new images behave very similarly to images that were forgotten. Unfortunately, due to the particularly high memory performance exhibited by our control subjects, few forgotten trials occurred. Thus, additional participants are required to further support this claim.
Chapter 4: Saliency Modeling

4.1: Overview

After completing the fixation and visual exploration analysis discussed in Chapter 3, we decided to implement a computational saliency model to further explore the data generated by the FVM task. The model we incorporated into our analysis was heavily inspired by the model utilized by Wang et al, 2015 and was adapted to received data from our FVM task. Our model consists of a linear support vector machine (SVM) classifier trained using fixation density maps extracted from the raw eye tracking data. When training our model, we looked at several different sets of features. The features we chose to include were semantic-level, object-level and pixel-level features but also introduced features for the image center and image background. We explored various subsets of the listed features in our testing to determine the model’s accuracy and competency. The directive of the model was to train the SVM using the selected subset of feature maps and fixation density maps to generate saliency weights for each feature. The SVM is essentially quantifying the contribution of each feature to the visualization of the images by taking a linear combination of the feature maps to best fit the fixation density maps. We individually trained the model per subject, meaning unique saliency weights were generated for each of our 19 individual participants. Further, we were able to split our data set by the number of presentations of the image. The model then output unique saliency weights for each subject for the two groups: new and repeated images. This allowed us to statistically compare the saliency weights across multiple image presentations. We did not investigate remembered versus forgotten models due to a limited number of trials that were forgotten and instead leave that for our future works.

4.2: Methods

The feature maps utilized within our model were provided alongside the OSIE dataset via Wang et al, 2015. To quickly summarize, pixel-level features were generated using the well-known Itti-Koch (Itti et al., 1998) saliency model. Object and semantic feature maps were generated by placing 2D gaussian blobs (\( \sigma = 2^2 \)) at the center of each object where the magnitude of the gaussian was the calculated object
value or manually labeled semantic feature value. Additionally, we generated feature maps for the image background by superimposing all feature maps for a given image and taking the negative space, i.e., not feature space, as the background. We generated image center feature maps similarly to semantic and object feature maps by placing a 2D gaussian blob ($\sigma = 1^°$) at the center of the image. All feature maps and stimulus images were downsized from 1600x1280 to 200x150 resolution to reduce model complexity and runtime.

The primary input to the computation saliency model was the human fixation density maps extracted from the raw eye tracking data collected during the FVM task. We first extract the fixations from the raw gaze data as seen in Figure 4.1 to isolate the fixations. The density maps were then created by performing a 2-D convolution on the isolated fixation maps. The kernel parameters of the convolution were the first factors we investigated to increase the performance of our SVM. In particular, we looked at the effect of altering both the patch size and the standard deviation of the kernel, an illustration of this parameter change can be seen below in Figure 4.2. Reducing the standard deviation of the kernel (Figure 4.2, left vs mid) provides less padding around the hotspots of the fixation density map, while reducing the patch size (Figure 4.2, left vs right) increases the resolution the maps but decreases the padding to provided maps more localized to individual fixations.

![Figure 4.1 Preprocessing of fixation map (left) and isolated fixations (right)](image)

Reducing the standard deviation of the kernel (Figure 4.2, left vs mid) provides less padding around the hotspots of the fixation density map, while reducing the patch size (Figure 4.2, left vs right) increases the resolution the maps but decreases the padding to provided maps more localized to individual fixations.
The remaining parameters that we experimented with were contained within the computational model itself, specifically within the SVM. The model determines the best fit for the saliency weights by using a linear SVM classifier to get the contribution of each included feature map for the recreation of the fixation density map. The SVM is essentially quantifying each features contribution to the visualization of the images by taking a linear combination of the feature maps to best fit the fixation density maps. The resulting quantification is the saliency weights which were calculated individually meaning each participant received unique saliency weights. To best determine the model’s performance, we incorporated a 5-fold cross validation where 80% of the data was used to train the model and the other 20% was used to test the model. The model performs these calculations by first taking a positive and negative sampling from the most and least salient regions of the training fixation density maps, respectively. This provided 4 additional parameters to consider: the number of positive samples, the number of negative samples, the top-salient cutoff percentage and the bottom-salient cutoff percentage. Unlike the preprocessing parameters, these model parameters are more difficult to visualize. Therefore, we look to the model's area under the Receiver Operating Curve (AUC) and the model’s testing accuracy as quantification of model performance.

During the journey to optimize our computational saliency model, we performed several iterations of parameter testing. Details of all performed parameter testing is described in detail in Table 1. First, we started by holding the kernel size constantly at
30 and allowed the standard deviations to vary (Table 1, Test 1, 2, 3). Using larger std. values lead to a very small increase in performance but recall from Figure 4.2, larger

Table 1: Compilation of parameter tests used to gauge the performance of our model on new vs repeated images. Note tests highlighted in green had the highest AUC and testing accuracy and tests weren’t necessarily completed in the order listed.

<table>
<thead>
<tr>
<th>Test</th>
<th>Kernel Size</th>
<th>Std.</th>
<th>Positive Pts</th>
<th>Negative Pts</th>
<th>Positive %</th>
<th>Negative %</th>
<th>New</th>
<th>Repeated</th>
<th>Avg Testing Accuracy</th>
<th>Avg AUC</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>10</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>80.9</td>
<td>80.52</td>
<td>0.8231</td>
<td>0.815</td>
<td>Semantic+B/C</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>15</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>81.33</td>
<td>80.92</td>
<td>0.8318</td>
<td>0.8254</td>
<td>Semantic+B/C</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>20</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>81.25</td>
<td>80.9</td>
<td>0.8318</td>
<td>0.8247</td>
<td>Semantic+B/C</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>30</td>
<td>40</td>
<td>83.75</td>
<td>83.54</td>
<td>0.825</td>
<td>0.8149</td>
<td>Semantic+B/C</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>15</td>
<td>75</td>
<td>375</td>
<td>30</td>
<td>40</td>
<td>82.8</td>
<td>82.83</td>
<td>0.8575</td>
<td>0.8518</td>
<td>ALL</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>5</td>
<td>20</td>
<td>80.78</td>
<td>80.78</td>
<td>0.8883</td>
<td>0.8873</td>
<td>ALL</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>5</td>
<td>40</td>
<td>88.29</td>
<td>87.94</td>
<td>0.9244</td>
<td>0.9206</td>
<td>ALL</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>86.67</td>
<td>86.65</td>
<td>0.9063</td>
<td>0.9053</td>
<td>ALL</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>10</td>
<td>100</td>
<td>300</td>
<td>10</td>
<td>30</td>
<td>83.29</td>
<td>83.32</td>
<td>0.8907</td>
<td>0.8907</td>
<td>ALL</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>84.55</td>
<td>84.25</td>
<td>0.8669</td>
<td>0.8616</td>
<td>Semantic+B/C</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>5</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>72.11</td>
<td>71.7</td>
<td>0.7937</td>
<td>0.7842</td>
<td>ALL</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>83.39</td>
<td>82.94</td>
<td>0.885</td>
<td>0.8831</td>
<td>ALL</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
<td>15</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>81.46</td>
<td>81.17</td>
<td>0.8577</td>
<td>0.8516</td>
<td>ALL</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>5</td>
<td>40</td>
<td>88.62</td>
<td>88.24</td>
<td>0.9103</td>
<td>0.9043</td>
<td>Semantic+B/C</td>
</tr>
</tbody>
</table>
std. makes the density map appear increasingly noisy. This led us to restrict std. values to, in all cases other than test 3, less than 50% of the patch size. Tests 10, 11 and 12 illustrate smaller patch sizes with std. of 50% patch size, these tests are comparable to test 8. Test 11 made it especially apparent that smaller patch sizes did not necessarily provide better performance. Next, we began increasing and decreasing the number of positive and negative sampled points. Initially, we decreased the positive points from

Table 2: Compilation of parameter tests used to gauge model performance on 1st, 2nd and 3rd image viewings. Note tests highlighted in green had the highest AUC and testing accuracy and tests weren’t necessarily completed in the order listed.

<table>
<thead>
<tr>
<th>Test</th>
<th>Kernel Size</th>
<th>Std.</th>
<th>Positive Pts</th>
<th>Negative Pts</th>
<th>Positive %</th>
<th>Negative %</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>Average Testing Accuracy</th>
<th>Average AUC</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>10</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>80.91</td>
<td>80.46</td>
<td>80.24</td>
<td>0.8229</td>
<td>0.8151</td>
<td>0.8068</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>15</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>81.39</td>
<td>80.87</td>
<td>80.75</td>
<td>0.8326</td>
<td>0.8247</td>
<td>0.8195</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>20</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>81.32</td>
<td>80.81</td>
<td>80.79</td>
<td>0.8321</td>
<td>0.8259</td>
<td>0.8209</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>30</td>
<td>40</td>
<td>83.64</td>
<td>83.62</td>
<td>83.12</td>
<td>0.8245</td>
<td>0.8159</td>
<td>0.8076</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>15</td>
<td>75</td>
<td>375</td>
<td>30</td>
<td>40</td>
<td>82.69</td>
<td>83.02</td>
<td>83.27</td>
<td>0.8581</td>
<td>0.8526</td>
<td>0.8399</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>5</td>
<td>20</td>
<td>80.84</td>
<td>80.76</td>
<td>81.01</td>
<td>0.8884</td>
<td>0.8856</td>
<td>0.8753</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>5</td>
<td>40</td>
<td>88.32</td>
<td>88.14</td>
<td>87.56</td>
<td>0.9248</td>
<td>0.9199</td>
<td>0.9067</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>86.7</td>
<td>86.66</td>
<td>86.37</td>
<td>0.9073</td>
<td>0.9047</td>
<td>0.8934</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>10</td>
<td>100</td>
<td>300</td>
<td>10</td>
<td>30</td>
<td>83.28</td>
<td>83.16</td>
<td>83.03</td>
<td>0.8902</td>
<td>0.8892</td>
<td>0.8799</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>84.58</td>
<td>84.39</td>
<td>83.81</td>
<td>0.8664</td>
<td>0.8616</td>
<td>0.8508</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>5</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>83.24</td>
<td>82.88</td>
<td>82.85</td>
<td>0.8851</td>
<td>0.8813</td>
<td>0.8705</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>10</td>
<td>75</td>
<td>375</td>
<td>10</td>
<td>40</td>
<td>83.24</td>
<td>82.88</td>
<td>82.85</td>
<td>0.8851</td>
<td>0.8813</td>
<td>0.8705</td>
</tr>
<tr>
<td>13</td>
<td>30</td>
<td>15</td>
<td>100</td>
<td>300</td>
<td>30</td>
<td>40</td>
<td>81.44</td>
<td>81.15</td>
<td>80.9</td>
<td>0.8584</td>
<td>0.8514</td>
<td>0.8395</td>
</tr>
</tbody>
</table>

Semantic+B/C

ALL

Semantic+B/C

ALL

Semantic+B/C

ALL
100 to 75 and increased the negative points 300 to 375, a net 25% change to both. Test 4 mirrors test 1 with the only difference being in the point sampling. From these tests we saw an increase in performance with fewer positive point and more negative points sampled. In theory, we could’ve incorporated more tests with varying sampling rates but we leave that for future works. Due to the nature of the fixation density maps, decreasing the cutoff percentage for negative sampling appeared to have a negative effect on the accuracy and AUC. This comparison can be seen between test 6 and 7 which prompted us to select 40% has the cutoff for our negative samples. Then, we rationalized that decreasing the cutoff percentage for positive sampling could further increase the model’s performance. To isolate the effect of the positive cutoff percentage, we performed test 8. Test 8 was identical to test 7 but with double the positive cutoff percentage, but increasing the cutoff actually led to lower model performance. The final parameter we explored was the set of feature maps used by the model. We compared models tested with all features included and then models trained with only the 12 semantic features plus the center and background. Tests 7 and 14 and also tests 2 and 13 are direct comparisons between feature sets, with models trained on all features outperforming semantic with background and center feature models in both

![Figure 4.3](image)

**Figure 4.3** Difference in saliency weights (left) generated by Test 7 and Test 14 for new vs repeated images alongside relative difference in weights (right)

comparisons. **Figure 4.3** showcases the difference between test 7 and test 14 for all semantic features and also background and center features. Notice the difference
between weights is very small in the left panel of Figure 4.3. The relative differences (right panel) fluctuate more in features, such as background and motion, that have comparatively small weights in both models. Due to the difference in weights being very similar, we selected the model trained on all features (test 7) as our primary model because it also had relatively higher AUC and accuracies compared to the model trained on only semantic features with background and center.

Additionally, most tests described in Table 1 were also performed on the ungrouped counterpart to new vs repeated images, where we did not pool the 2nd and 3rd viewings into repeated. The results of the parameter tests in Table 2 displayed very similarly to those in Table 1. Ultimately, the best performance observed was during test 7 by taking 75 pixels from the 5% most fixated regions and 375 pixels from the 40% least fixated regions while having a kernel size of 30 px and std. of 10 px. This set of parameters was used in all saliency and model analysis going forward.

4.3: Results / Discussion

The goal of Chapter 4 was to modify a computational model to receive our human fixation density maps as inputs and output saliency weights by feature. We also wanted to tune our model to provide the most optimal performance, which required extensive parameter testing. After determining the most optimal parameters to use within our saliency model, we had to generate our saliency weights. Initially, we determined the saliency weights by averaging feature weights generated by the 5-cross fold validation that the model performs but we also decided to compare the average weights with the weights of the best performing model seen in Figure 4.4. For clarity, the best performing model was the model that had highest overall testing accuracy. Graphically, the two methods seem to produce very similar weights.
To be sure, we looked at the relative difference in the weights generated by the two methods in Figure 4.5. The relative difference was calculated as the absolute difference between the two weights divided by the largest observed weight for that feature. For most features the relative difference is very small, less than 20% for most weights. The large relative difference seen in the 3rd viewings of sound and smell is due to the fact that the magnitude of those weights, even the maximum of the observed weights, is
very small causing unexpected side effects. As a result of this simple test, we decided to select the averaged weights instead of the best performing as our method for condensing the 5 weights generated by 5-fold cross validation.

Now that we have discussed all the methods required for the model to generate saliency weights, we explore the weights of the best performing model. The weights generated by the best performing model are showcased in Figure 4.6, the left panel contains all weights while the right contains all features excluding the center feature. The center feature is excluded from the right panel solely for readability, as its weights are almost an entire order of magnitude larger than the other features.

This suggests that the center feature contributes the most to the model’s depiction of the fixation density maps. This large difference in magnitude could also be artificial noise generated by the nature of the FVM task. After each button-press, the participants’ gaze is relatively centered due to the presentation of the on-screen button box and a mid-screen cross that appears briefly. Thus, the magnitude of the center feature may be unintentionally inflated by the task. From Figure 4.6 we can see that the saliency weights differ between new and repeated images for the background, color, orientation, gazed, text, watchability and operability features. Orientation, size, complexity and
eccentricity are the most dominating of the pixel and object level features with eccentricity having a relatively large negative saliency weight. A closer look at the semantic weights shows face and emotion are the most dominating semantic features suggesting these features contributed more heavily to the recreation of the fixation density maps than other semantic features. Interestingly, the saliency weights for gazed, text and watchability trend significantly lower in repeated images suggesting that those features are more important to the recreation of the fixation density maps in initial viewings than they are in repeated viewings. On the other hand, operability weights are significantly higher in repeated images suggesting that the operability feature is more important in repeated viewings than in the initial viewing.

![Saliency Weights Semantic Features Only](image)

*Figure 4.7 Saliency weights generated by the best performing model on 1st, 2nd and 3rd viewings. (t-test, ***p<0.001, ** p<0.01, * p<0.05, error bars denote SEM across subjects)*

For completeness and a more in-depth look, we also performed the same analysis of the previous section on 1st, 2nd and 3rd viewings. All model parameters were kept as those from the best performing model, the only difference was that 2nd and 3rd viewings were now considered distinct instead of being grouped into the repeated category. Initially, we pooled the 2nd and 3rd fixations into the repeated group because we thought it would be best for the saliency model to have additional images for training (only 50 images were viewed three time providing the model 40 images to train on and 10 images to test on). But as seen in Table 2, the model can indeed operate on smaller
image sets of this size without substantial loss of accuracy. The saliency weights generated from this analysis are depicted in Figure 4.7, note once again the center feature was excluded from this figure because of the excessive magnitude difference and no observed difference between repetitions. Table 3 is included to highlight the differences observed between 1\textsuperscript{st}, 2\textsuperscript{nd} and 3\textsuperscript{rd} viewings in a more concise manner.

As in the new vs repeated comparison, significant differences were also observed in the background, color, orientation, gazed, text, watchability and operability features. Additionally, differences were observed in complexity, convexity, touched and touch features that were not observed in the new vs repeated comparison.

Unsurprisingly, the differences in these 4 features were all observed between the 1\textsuperscript{st} vs 2\textsuperscript{nd} viewing and all but touch displayed differences between the 2\textsuperscript{nd} and 3\textsuperscript{rd} viewings. It’s also interesting to note that of all 7 features with observed differences in new vs repeated comparisons, only two (color and operability) features did not display differences in the 1 vs 2\textsuperscript{nd} viewing. Conducting the analysis on the 1\textsuperscript{st}, 2\textsuperscript{nd} and 3\textsuperscript{rd} viewings gave us an additional, more comprehensive look at how feature saliency is affected by repeatedly viewing the same image.

<table>
<thead>
<tr>
<th>Feature</th>
<th>1st vs 2nd</th>
<th>1st vs 3rd</th>
<th>2nd vs 3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>***</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>Color</td>
<td>n.s.</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>Orientation</td>
<td>***</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td>Complexity</td>
<td>n.s.</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Convexity</td>
<td>n.s.</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td>Eccentricity</td>
<td>n.s.</td>
<td>**</td>
<td>*</td>
</tr>
<tr>
<td>Touched</td>
<td>n.s.</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>Gazed</td>
<td>*</td>
<td>n.s.</td>
<td>*</td>
</tr>
<tr>
<td>Touch</td>
<td>n.s.</td>
<td>**</td>
<td>n.s.</td>
</tr>
<tr>
<td>Text</td>
<td>*</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td>Watchability</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Operability</td>
<td>n.s.</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Chapter 5: Viewing Consistency

5.1: Overview

After completing the computational saliency modeling discussed in Chapter 4, we decided to further investigate the consistency of viewing between image presentations. In addition to the fixation analysis discussed in previous chapters, we further explore the fixations. More precisely, we explored whether the fixations after the first saccade land on the same objects (note we were able to analyze whether a fixation landed on the same category of objects as in (de Haas et al., 2019) but additional refinement of our framework could allow us to analyze instead by identical objects). We quantified the viewing consistency (VC) using two methods: ordered fixations and total image. The difference between the two is discussed in detail in the next section. Simply put, we consider each fixation individually in the sequence that they occur and also how the image is viewed as a whole. Also, we examined how the viewing consistency is affected by the presence of semantic features in image subsets as we did for the behavioral data analysis in Chapter 2. Additional secondary analysis was performed to investigate how often semantic features were viewed exclusively in novel images (1st viewing) but not in repeated (2nd or 3rd viewing) images and vice versa. Finally, we finish this section with an in-depth look at a number of correlation analyses performed on human fixation density maps.

5.2: Methods

The objective of this section is to analyze the viewing consistency of our FVM task. In order to do so, first we overlaid the fixations from both the novel and repeated images onto the stimulus image. No trials were excluded. To include all trials, the number of fixations were capped to be the minimum number of fixations from the new (or novel) and repeated viewings. This allowed us to do a one-to-one comparison between the ordered fixations. Note images that were viewed three times were pooled into the repeated group together with images viewed twice. Figure 5.1 illustrates the
different patterns in fixations for the novel and repeated groups (fixations are colored green and blue, respectively) with semantic feature objects outlined in yellow.

![Figure 5.1 Ordered fixations from novel and repeated image viewing with semantic objects highlighted in yellow, novel fixations in green and repeated fixations in blue.](image)

The first metric we explore in this section is whether the ordered fixations landed on the same features. To quantify this metric, we calculated the number of trials where the same category of feature is fixated on in both the novel and repeated viewings. Fixations are analyzed independently, so we calculated the number of trials where the first fixation lands on the same category of feature in both the novel and repeated viewings. We replicated this process for all numbered fixations up to 10 fixations. The logic behind capping the number of fixations came from the percentage of trials that actually display that quantity of fixations. As the number of fixations increased the percentage of usable trials decreased as seen in Table 4. The percentage of trials containing 10 or more fixations dropped-off quite rapidly, which is precisely why we chose to cap the fixations to 10 for this analysis. The average viewing consistency was found in each person by summing the number of consistent trials and dividing by the
total number of trials. The viewing consistencies were then averaged across subjects to produce the VC figures shown in the following results section.

**Table 4 Breakdown of usable trials for each quantity of fixations selected**

<table>
<thead>
<tr>
<th>Number of Fixations</th>
<th>Number of Usable Trials</th>
<th>Percent of Total Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2827</td>
<td>99.19%</td>
</tr>
<tr>
<td>2</td>
<td>2819</td>
<td>98.91%</td>
</tr>
<tr>
<td>3</td>
<td>2801</td>
<td>98.28%</td>
</tr>
<tr>
<td>4</td>
<td>2774</td>
<td>97.33%</td>
</tr>
<tr>
<td>5</td>
<td>2687</td>
<td>94.28%</td>
</tr>
<tr>
<td>6</td>
<td>2570</td>
<td>90.18%</td>
</tr>
<tr>
<td>7</td>
<td>2355</td>
<td>82.63%</td>
</tr>
<tr>
<td>8</td>
<td>1986</td>
<td>69.68%</td>
</tr>
<tr>
<td>9</td>
<td>1530</td>
<td>53.68%</td>
</tr>
<tr>
<td>10</td>
<td>986</td>
<td>34.60%</td>
</tr>
</tbody>
</table>

In the previous section, we analyzed the images by ordering the fixations and performing comparisons on a fixation-by-fixation basis. In this section, we took a quick look at the viewing consistency for the entire image presentation by pooling all the fixations. This allowed us to get a more generalized quantification of the viewing consistency on a per image basis. For each visualization, we analyze the feature category for all fixations in the viewing. No fixations were excluded and the fixations were not capped to 10 as in the previous section. The list of features viewed was documented for the 1st, 2nd and 3rd viewings and the features lists were compared accordingly. We reuse our definition of viewing consistency as before: the ratio of trials where the exact same feature categories were viewed. This method provided us three comparisons to perform: 1st vs 2nd, 1st vs 3rd and 2nd vs 3rd.

Next, we revisit both the ordered fixations and the total image viewing consistency but now we reapply those methods on the semantic feature subsets as we described in Chapter 2. The methods for constructing the subsets are exactly the same.
as previously described but a short summary is provided here. We construct a subset for each feature by finding all the images in our set that contain a given semantic feature and then repeat this process for all 12 semantic features. In this analysis, we did not analyze the viewing consistency of the complement subsets (those not containing a given feature) and instead leave that for our future works.

The next topic explored was how often semantic features were gazed at in either exclusively in the novel image or in the repeated image. Essentially this means how often do features get looked in the initial viewing and ignored in latter viewings but also how often did features get ignored until later viewings. To quantify this, we counted the number of instances where a feature was viewed in the first viewing but not in the second viewing and also viewed in first viewing but not in the third viewing. Also, the opposite cases, in the second viewing but not in the first and in the third viewing but not in the first, were considered. In this analysis, we did not compare 2nd vs 3rd and vice versa. All comparisons made here were normalized by the total number of viewed images that contained the specific feature.

The final avenue explored in Chapter 5’s analysis was the correlation of human fixation density maps. The comparisons made here were between the fixation density maps of the repeatedly viewed images. The correlation coefficients were calculated using MATLAB built in “corr” function which takes two vectorized images as parameters. First, we looked at the comparisons between 1st vs 2nd, 1st vs 3rd and 2nd vs 3rd. To calculate the correlation, we input the vectorized images into the “corr” function and received a “r” value for each comparison. We then averaged all correlation coefficients to get a correlation score per individual, that score is then averaged across subjects to get the final average correlation coefficient. This correlation method was also implemented on images subset by semantic feature as discussed in detail previously. Additionally, we include a look at the correlation comparison of images with more and less than average objects. Finally, this correlation method was used to develop several different comparisons including those for a shuffled label test. The shuffled label tests were done within a presentation to determine if innate visualization patterns existed in our images. To perform this shuffled label test, we simply created a list of all images (by
image id number) that the participant viewed and then made a second list by taking a random permutation of the first list. Then we performed our correlation analysis between the unshuffled list images and the shuffled list images.

5.3: Results / Discussion

5.3.1 Viewing Consistency

The first viewing consistency metric we explored was whether ordered fixations landed on the same features. We quantified this metric by analyzing the number of trials where the exact same feature categories were fixated on in both the novel and repeated viewings. This one-to-one comparison between fixations allowed us to generate Figure 5.2 which illustrates the ratio of trials where similar features were viewed. Note that as we incremented through fixations the ratio of similar trials decreased and also the magnitude of the standard error increased. This increase is due to the drastic reduction of usable trials with more than 10 fixations as discussed previously in Table 4. The decrease in viewing consistency was likely caused by a more focused attention in repeated image viewings. As we have shown in Chapter 3, the number of fixations and dispersion area decreased while the fixation duration increased with repeated viewings.

Figure 5.2 Ratio of trials where ordered fixations occupied the same feature categories in both novel and repeated images. (Error bars denote SEM across subjects)
We believe that this focused exploration is a contributing factor to the trend we saw in the viewing consistency of ordered fixations.

To further explore the viewing consistency of ordered fixations, we revisited the different feature subsets of our image set. As previously discussed, we investigated the VC ratio for only the images that contained our twelve semantic features. The results of this investigation are depicted above in Figure 5.3. For most cases, the VC ratio displayed a trend similar to the ratios taken from the entire image set, starting relatively high in the initial fixations but decreasing quite rapidly afterwards, leveling out at around approximately a ratio of 50%. From these results, we decided to disregard the order of the fixations to explore a more generalized concept of viewing consistency.

Figure 5.3 Ratio of trials where ordered fixations occupied the same feature categories in both the novel and repeated viewings of images that contained the given semantic feature. (Error bars denote SEM across subjects)
In the previous section, we analyzed the images by ordering the fixations and performing comparisons on a fixation-by-fixation basis. This ordered analysis provided us useful information on viewing consistency trends but this definition of viewing consistency did not necessarily provide a useful means of comparison. In this section, we instead looked at the viewing consistency for the entire image presentation by pooling all fixations and considering all repetitions individually instead of grouping the 2nd and 3rd viewings. This allowed us to get a more generalized quantification of the viewing consistency on a per image basis. Figure 5.4 shows the results of this generalized comparison detailing the viewing consistency ratios for 1st vs 2nd, 1st vs 3rd and 2nd vs 3rd viewings. While all three ratios tended to be quite high, the 1st vs 3rd ratios on average tend to be significantly (t-test, p<0.05) greater than 1st vs 2nd viewings. This suggests that in terms of features viewed, the 1st viewing of an image is on average, more similar to the 3rd viewing than it is to the 2nd viewing.

To continue this analysis, we further split the image set into the semantic feature subsets as discussed in detail in **Chapter 2**. Reperforming the pooled viewing
consistency analysis on the semantic image sets resulted in Figure 5.5. Many features, touched, gazed, motion, sound, smell, touch and taste, surprisingly did not exhibit any significant differences in viewing consistency between image viewings. On the other hand, differences were seen between 1\textsuperscript{st} vs 3\textsuperscript{rd} but also between 1\textsuperscript{st} vs 2\textsuperscript{nd} and 2\textsuperscript{nd} vs 3\textsuperscript{rd} for face, taste, watchability and operability feature sets. The levels of significance did vary between the four features, see Figure 5.5, but the trend was the same: the VC ratio for 1\textsuperscript{st} vs 2\textsuperscript{nd} was significantly higher than the ratios for 1\textsuperscript{st} vs 3\textsuperscript{rd} and 2\textsuperscript{nd} vs 3\textsuperscript{rd}. This result suggests that in images containing these semantic features, the first and second presentations are the most similar viewings. This trend is also observed for the emotion.
feature sets except only between the 1st vs 2nd and 1st vs 3rd comparisons. It should also be noted that when compared to the VC ratios of the total image set, these feature subsets have substantially lower ratios in all categories.

Another interesting metric explored in this section was how often these semantic features were viewed exclusively in initial viewings but not in repeated viewings and vice versa. Essentially, we get a normalized count of the number of feature objects that were fixated on in one viewing but not the other. We first make this comparison between the 1st and 2nd viewings in Figure 5.6 (note blue denotes object viewed exclusively and in the first view, as does red for the second view). Unsurprisingly, the number of new objects were relatively similar for most features suggesting that the likelihood of viewing new semantic objects is relatively the same across viewings. Text feature had an increased likelihood to be viewed in the second viewing opposed to the first viewing. This comparison process was repeated for the 1st vs the 3rd and vice versa as seen in

![Figure 5.7](image_url)

*Figure 5.7. Unfortunately, this comparison is much noisier than the previous comparison. Specifically, the error bars for some features are much greater in*
magnitude for the 3rd viewing than either the 1st or 2nd viewing, likely due to the decreased number of images (200 trials in 1st, 100 trials in 2nd and 50 trials in 3rd) viewed three times. In turn, we do not see much difference in the occurrences of features viewed exclusive in the 1st viewing and not the 3rd and vice versa.

Figure 5.7 Average number of trials where a fixation occurred exclusively in the first or third viewing normalized by feature count. (Error bars denote SEM across subjects)

5.3.2 Correlation Analysis

Another metric we explored in this chapter was the correlation between fixation density maps of repeated viewings. First, we calculated the average correlation per subject by averaging the correlation coefficient between density maps for each of the three cases shown in Figure 5.8, before averaging across subjects. Figure 5.8 indicates that the correlation coefficient between first vs second viewings is higher than either first vs third viewing or second vs third viewing correlations. This trend is
very similar to the trend exhibited in semantic features with significant differences in VC, seen in Figure 5.5. In addition, the second vs third correlations were also significantly higher than the first vs third correlations. This suggests that second and thirds viewing behave more similarly than first and third viewings but not as similarly as first and second viewings. We expect that some similarity occurs instinctively due to the nature of the images. To combat this theory, we examined the effect of randomly shuffling the maps for each case and recalculating the correlation coefficients. Note that only the associated labels were shuffled, the fixation density maps were not altered for this correlation comparison. For each case in Figure 5.9, the average correlation coefficients were statistically and substantially lower when computing with shuffled maps. This result suggests that the correlations observed in Figure 5.8 are not a result of random image patterns but instead are natural patterns exhibited during visualization.
Figure 5.9 Correlation analysis of fixation density maps with ordered vs shuffled labels for all images. (*** $p < 0.001$, error bars denote SEM across subjects)

Figure 5.10 Correlation Comparison between repeated viewings of images with greater than or less than average objects (* $p < 0.05$, ** $p < 0.01$, error bars denote SEM across subjects)
To wrap up our correlation analysis, we shift to the feature image subsets that do and do not contain semantic features. This provided us a total of 12 sets of correlation comparisons which are summarized in Table 5. Additionally, we included a comparison for objects as discussed previously in Chapter 2. In short, we divide our image set into images with more than average number of objects and less than average objects.

To perform the comparisons, we calculate the average correlation coefficient between the numbered fixation density maps for the feature and non-feature image sets. Note that all completed comparisons differed significantly from their shuffled counterparts in all three of our tested categories at p < 0.001 significance. The result of the object comparison is included in Figure 5.10 where the correlation coefficients for the more object set were significantly higher than the less object set for each case. We replicate these comparison methods for the feature vs non-feature image sets and tabulated the results in Table 5. Unlike the object comparison, only a few differences were observed in the feature comparisons for the emotion, face, gazed, sound, touch and touched feature sets.

Table 5 Mean correlation coefficient between all semantic subsets and their corresponding no-feature subset. ( ***p<0.001, ** p<0.01, * p<0.05)

<table>
<thead>
<tr>
<th>Feature</th>
<th>1st vs 2nd</th>
<th>1st vs 3rd</th>
<th>2nd vs 3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion</td>
<td>***</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Face</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Gazed</td>
<td>n.s.</td>
<td>n.s.</td>
<td>*</td>
</tr>
<tr>
<td>Motion</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Operability</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Smell</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Sound</td>
<td>*</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Taste</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Text</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Touch</td>
<td>**</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td>Touched</td>
<td>**</td>
<td>n.s.</td>
<td>**</td>
</tr>
<tr>
<td>Watchability</td>
<td>n.s.</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
</tbody>
</table>
Chapter 6: Comparison of Visual Attention in ASD

6.1: Overview

The penultimate chapter of this thesis focuses on implementing a selection of the previously described and defined methods for participants diagnosed with autism spectrum disorder and comparing with the control results. The remaining four sections of this chapter correspond with the previous four chapters of Behavior and Memory, Fixation and Gaze Exploration and Viewing Consistency. Unless otherwise stated, the analysis on those with ASD was performed exactly as it was in the analysis of the neurotypical individuals. Our lab recruited 9 persons with ASD to participate in our FVM task with a breakdown of 7 males and 2 females. Unfortunately, data collected from 1 of the female participants was excluded due to the loss of an entire session of data. This resulted in 7 males and 1 female (mean age $21.57 \pm 5.91$) participants with usable eye tracking and behavioral data. The following sections will illustrate the comparisons of methods on those with ASD and neurotypicals.

6.2: Behavior and Memory

From the 16 usable sessions of ASD data, fixation data was extracted and sorted into their corresponding memory events. We first show in Figure 6.1 that ASD participants memorized our images well with high average hit and correct rejection rates comparable to the performance of controls. Note once again the vastly lower percentage of fixations for the incorrect response types of miss and false alarm. Unlike in the control analysis, differences were not observed between hit and miss rates or correct rejection and false alarm rates. Note unlike the controls both the repeated (Hit and Miss) and new (Correct Rejection and False Alarm) ASD trial types, higher fixation durations occurred during incorrect response trials than in the corresponding correct response trials. While the differences were not significant (t-test $p = 0.441$, $p = 0.603$) this suggests that during incorrect response trials, those with ASD might gaze at fewer locations but fixate on those locations longer than they would in correct response trials.
As with controls, no notable trends were revealed by the reaction time for those with ASD given that all four memory trials had very comparable reaction times. Also, the confidence ratings did not differ significantly across the four memory trials but subjects show very high overall confidence in their selections. Average confidence ranged from $1.049 \pm 0.016$ to $1.100 \pm 0.06$, implying that participants selected a “very sure” confidence response for a majority of trials.
Next, we directly compare the total image behavior between the control and ASD groups, seen in Figure 6.2. The fixation duration of both hit and correct rejection appears to be greater in controls than in ASD but their t-tests report $p = 0.063$ and $p = 0.117$, respectively. All other $p$-values were much larger than those required for 95% confidence. As a result, no significant differences were observed in any behavioral comparisons made between control and ASD groups.

6.3: Fixation and Gaze Exploration

The first comparison we explore here is the new and repeated images between controls and those with ASD. Applying all of the metric defined in the methods of section 3.2, provided the results in Figure 6.3. The number of fixations trend very similarly for both controls and ASD. The controls appear to have increased fixation durations, decreased fixation dispersion, and decreased center bias for new and repeated images. Unfortunately, these suggested trends are just an educated guess because all $p$-values were found to be greater than .05.

![Figure 6.3](image_url)  
*Figure 6.3 New and repeated Images for Control vs ASD: Number of Fixations (A), Average Fixation Duration (B), Fixation Dispersion (C) and Center Bias (D). (Error bars denote SEM across subjects)*

The last comparison we investigate in this section is between memorized and forgotten images for the ASD and control groups. Applying all of our previously defined metric to the sorted trials provides the results presented in Figure 6.4. None of our test
criterion proved to differ significantly for remembered and forgotten images between controls and those with ASD. The trends we see developing are very similar number of fixations, decreased fixation duration, increased fixation dispersion and increased center bias in those with ASD for both remembered and forgotten images. To further support this claim would require additional participants for both groups but especially those with ASD.

Figure 6.4 Remembered and Forgotten Images for Control vs ASD: Number of Fixations (A), Average Fixation Duration (B), Fixation Dispersion (C) and Center Bias (D). (Error bars denote SEM across subjects)

6.4: Saliency Modeling

Following the inconclusive results of sections 6.2 and 6.3, we implemented our computational model on the ASD data. The model we used to generate the saliency weights for the ASD group was identical to the most optimal model discussed thoroughly in Chapter 4. To implement this comparison between the two groups, we normalized the saliency weights within each subject. The largest weight was determined for each subject and all remaining weights were divided by the largest weight. This produce a weight of 1 for a given feature in each subject, typically that feature was the center feature. The normalized weights were then averaged across subjects to produce the final weights for new and repeated images to compare. Figure 6.5 illustrates the
saliency weights corresponding to new images while Figure 6.6 shows those for repeat images.

Many differences were observed in the saliency weight comparisons, those that were significantly different are tabulated in Table 6. Surprisingly, of all the features that exhibited differences in weights, the normalized ASD weights were significantly greater
than those of the control for all features except for the center feature. This discrepancy is particular interesting due to the fact that those with ASD tend to be more biased toward the center of images. Another interesting case is the face feature which was higher in both new (non-significant) and repeated images for those with ASD. We would expect those with ASD to exhibit increased avoidance to social stimulus but this was not the case in this analysis. The small sample size of 8 persons with ASD is likely a major contributing factor to the discrepancies seen in this result. Thus, we require additional participants to further explore the saliency weights of those with ASD.

### 6.5: Viewing Consistency

The objective of this final section was to explore if our viewing consistency metrics differed between control and ASD subjects. To accomplish this, we compared the ordered fixation and total image viewing consistency between the two groups. These metrics were applied to the ASD data in the exact same manner as the control data which was described in detail in Chapter 5. The ordered fixation comparison,
displayed in Figure 6.7, illustrates that the trend in viewing consistency of ordered fixations is very similar across groups. A significant difference is only seen in the 2\textsuperscript{nd} of the 10 fixations considered, where the VC ratio is significantly lower in those with ASD. Additionally, this trend is seen in the first fixation except the difference is not significant (p = 0.061), while the 3\textsuperscript{rd} and following fixation ratios have very large p-values. This suggests that between groups the viewing consistency is higher in controls for the initial fixations but tend to be quite similar for subsequent fixations.

Figure 6.7 Control vs ASD: Ratio of trials where ordered fixations occupied the same feature categories in both novel and repeat Image (* p < 0.05, error bars denote SEM across subjects)
Figure 6.8 Control vs ASD: Viewing consistency ratios for all images disregarding fixation order. (* p<0.05, error bars denote SEM across subjects)

The second metric we used to compare the groups was the total image viewing consistency seen in Figure 6.8. Considering the results of the ordered VC analysis, it was unsurprisingly that the viewing consistencies between the 1\textsuperscript{st}, 2\textsuperscript{nd} and 3\textsuperscript{rd} viewings were very similar in those with ASD compared to the controls. From these metrics we conclude that the images were viewed consistently by both groups regardless of presentation iteration.
Chapter 7: Conclusions and Future Direction

Conclusions:

This thesis contains a comprehensive analysis of behavior and visual exploration during our Free Viewing Memory task. The behavior section showed that subjects memorized images well with high rates of hit and correct rejection as opposed to low rates of miss and false alarm, respectively. The elevated rates of hit and correct rejection also extend to the duration of fixations with both having a significantly higher duration than their counterparts of miss and false alarm, respectively. In particular, these higher fixation durations in correct response trials, translated to participants gazing at fewer locations but focusing on these fixations for longer than they would in incorrect response trials. In addition, we devised an image sub-setting strategy to study the effects that particular features have on our behavior metrics. In particular, this strategy showcased differences in both emotion and face features. Images containing emotion objects corresponded to higher hit and correct rejections compare to those without, the opposite trend is exhibited by images containing face features.

Next, we devised the four metrics of fixation numbers, fixation duration, fixation dispersion area, and center bias to quantify the visual exploration during our task. Novel images contained more fixations with lower fixation durations, higher dispersion and a larger center bias when compared to images that were repeated. A comparison between remembered and forgotten images was less conclusive, due to limited forgotten trials, and remembered images contained less fixations than forgotten images. This outcome, together with data trends, suggests that forgotten image could behave similarly to novel images. Then, we adapted a computational model from Wang et al. 2015 to analyze fixation density maps. After performing extensive tuning of the model and input parameters, we achieve average accuracies as high as 88% with corresponding AUCs as high 0.924. This model calculated saliency weights for comparisons between new vs repeated images and also 1st vs 2nd vs 3rd image viewings. Results showed saliency weights differed in the background, color,
orientation, gazed, text, watchability and operability features between new and repeated images.

Next, we devised two different strategies to quantify view consistency, by ordered fixations and then generalized by total image. Ordered fixations, even in feature subsets, showcase a very general trend of high VC in early fixations that tapers off with increase fixations. In the generalized VC, differences in $1^{st}$ vs $2^{nd}$ and $1^{st}$ vs $3^{rd}$ suggests that the $1^{st}$ viewing is more similar to the $3^{rd}$ viewing than to the $2^{nd}$. In feature subsets, the total VC showed significant differences in face, emotion, taste, watchability and operability features while the likelihood of viewing new semantic objects stayed relative the same across viewings. Correlation analysis performed on the fixation density maps showed that similarity ranked in descending order as $1^{st}$ vs $2^{nd}$, $2^{nd}$ vs $3^{rd}$, then $1^{st}$ vs $3^{rd}$.

Finally, we repeated our previously discuss methods to analyze data collected from those with ASD and compared with controls. Ultimately, no significant differences were seen in our behavior metrics and none were observed in either the new vs repeated or remembered vs forgotten comparisons of our gaze exploration metrics. On the other hand, many differences were observed in the saliency weight comparison between controls and ASD, the full table of results included in Table 6. Also, generalized VC was very similar between groups with the notable exception of ordered fixations having higher VC ratios in controls for initial ordered fixations.

**Future Works:**

The results of this thesis could be extended further to provide an even more comprehensive analysis of attention and memory in visual exploration. Primarily, the best way to strengthen the results of this thesis is to acquire additional data from controls and especially from additional ASD participants. Alongside supplemental data, additional tuning of the computational model parameters could provide increased model performance. As well, it would be very interesting to independently train the model on ASD data to determine if differences occur in the reported saliency weights. Also, an in-depth analysis of ASD data, such as that performed in Chapters 2-5 for controls, could provide useful insight into ASD behaviors. Examining the social feature subsets
described in Chapters 2 and 5 could be especially beneficial. Other additions that could perhaps improve this analysis includes reworking the computational model to use more sophisticated classification methods such as a neural network in place of our linear SVM, designing more advanced metrics to quantify visual exploration and converting our viewing consistency methods to analyze by individual objects instead of by feature categories.
Chapter 8: References


Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, Qi Zhao; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7521-7531


